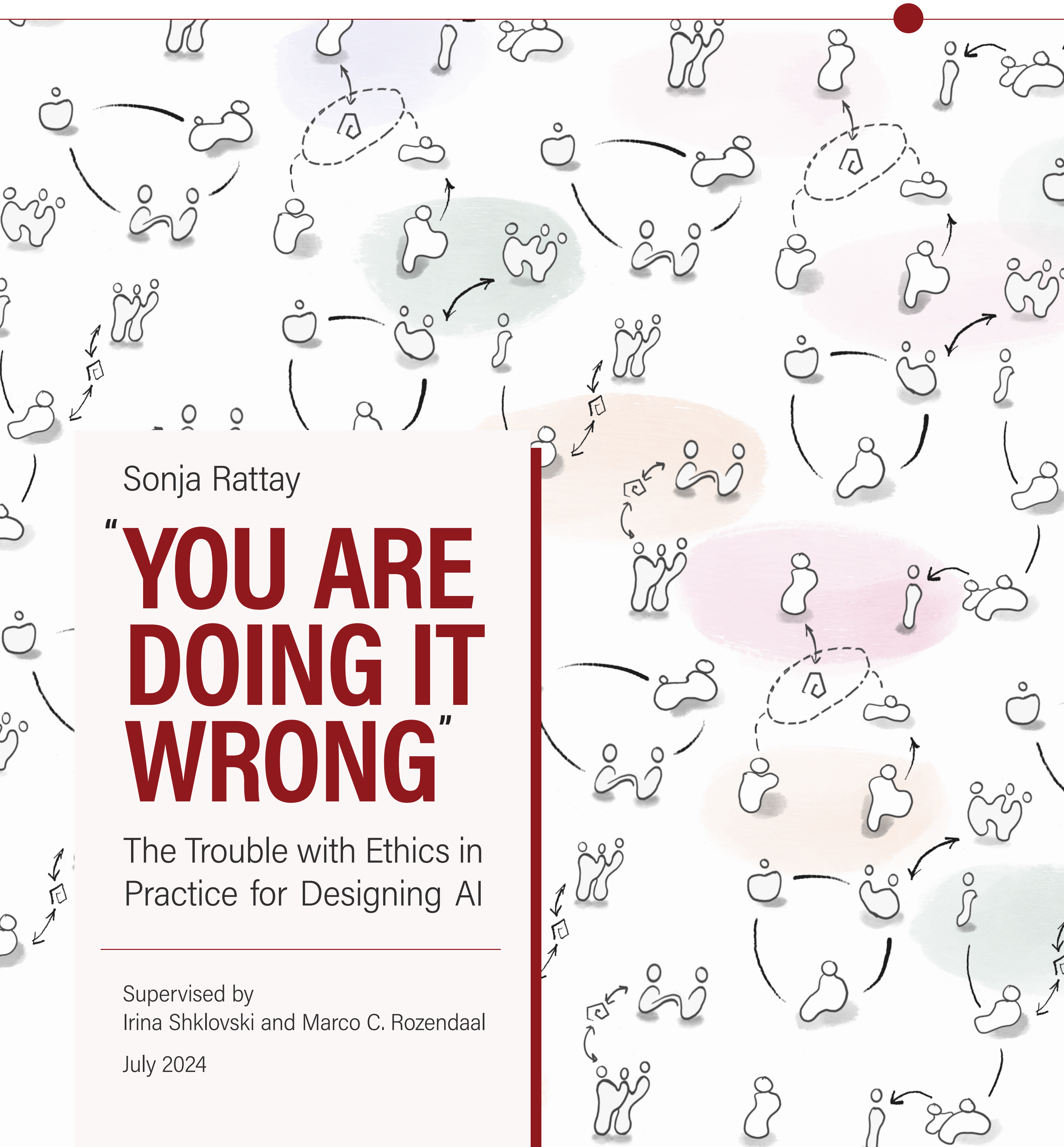
UNIVERSITY OF COPENHAGEN



Sonja Rattay

"YOU ARE DOING IT WRONG"

The Trouble with Ethics in Practice for Designing AI

Supervised by
Irina Shklovski and Marco C. Rozendaal

July 2024

Sonja Rattay
"*You are doing it wrong*" - *The Trouble with Ethics in Practice for Designing AI*
PhD in Human Centred Computing, July 2024
Supervisors: Irina Shklovski and Marco C. Rozendaal

University of Copenhagen
Faculty of Science
PhD in Human Centred Computing
Blegdamsvej 3B
2200 Copenhagen N

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen, July 14, 2024

This work is part of the DCODE project. The project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955990.

Abstract

Ethics in the design of technology has a long history of debate, that periodically surfaces with more urgency than at other times. The fast progression of development and deployment of AI systems in the last years is one of such times. As a result, conversations around ethics have spread beyond the field of moral philosophy, and the disciplines of HCI, computer science, data science and engineering are working towards creating generally applicable forms of applied ethics for the field. Of course, many controversies remain, and debates around what such applied ethics might look like touch on broad political and economical tensions, with plenty of critical studies highlighting the interconnectedness of technological progress and social issues.

In this thesis, I examine such tensions from the perspective of practice of designing technology, with particular consideration of relational aspects of ethics and the social dimensions informing the envisioning and making of technology. Ethics in this framing is considered as the ongoing process of making decisions based on moral judgements. My work takes this examination beyond the exploration of normative judgements and prescriptive guidelines, into the fields of envisioned futures, and affective experiences of moral tensions and value trade offs.

I present three main contributions:

First, I outline the productive influences of sociotechnical imaginaries on design practice and understandings of ethics tied to it. I analyse existing case studies of well intentioned design projects from the field of STS. I highlight three important dimensions in which such present sociotechnical imaginaries preconfigure how technicians might engage with value work - initiatives that aim to guide a more ethical conduct. I also provide a case study of such identification of present sociotechnical imaginaries in the context of AI and sustainability. Based on this case study, I present a framework and vocabulary, based on the philosophical concept of the device paradigm by Borgman, that helps to understand the seductive dynamics of such imaginaries and how it affects and shapes the relation between morality and technology for technicians. I argue that if underlying present sociotechnical imaginaries are not identified and clearly addressed, any following work on adhering to values or guidelines will be interpreted in favor of such existing imaginaries.

Second, I present a method for engaging and challenging existing sociotechnical imaginaries with practitioners and researchers of sensor networks. This method uses speculative roleplay, and centers value tensions through the lens of care ethics as a situated approach to responding to socially challenging situations, in which people encounter vulnerabilities and the contextual limitations of technology. This method presents a productive approach to challenging imaginaries by foregrounding situatedness, performativity, multiple voices and actors and interdependence of needs.

Third, I engage in ethnographic work with a team of practitioners over a longer time span to investigate the emotional aspects of integrating ethics toolkits into existing technical workflows. I apply the concept of moral stress, based on prior research in moral psychology, management studies and nursing studies, to the increasingly moralised work tasks of developers, and trace the impact of this experience through the everyday work experience of technical practice. This chapter also ties back into pinpointing where practitioners encounter sociotechnical imaginaries at play and the stress it causes to shift mindsets around them.

My main argument is that in order for any ethical interventions to be successful, we need to start with the envisionings of technology and its impact on social and political life, and can not stop with interventions to shift those, but need to also account for the affective fall out of such engagements experienced by the practitioners within the broader field of their practice.

Abstract

Etik I teknologidesign har en lang historie af debat, som periodisk dukker op med mere hastende behov end på andre tidspunkter. Den hurtige udvikling og implementering af AI-systemer i de seneste år er sådan et tidspunkt. Som følge heraf har samtaler om etik spredt sig ud over feltet af moralfilosofi, og disciplinerne HCI, datalogi, data science og ingeniørvidenskab arbejder på at skabe generelt anvendelige former for anvendt etik for feltet. Selvfølgelig fortsætter mange kontroverser, og debatterne om, hvordan sådan anvendt etik kunne se ud, berører brede politiske og økonomiske spændinger, med mange kritiske studier, der fremhæver sammenhængen mellem teknologisk fremgang og sociale anliggender.

I denne afhandling undersøger jeg nogle af disse spændinger fra et praksisperspektiv inden for design af teknologi, med særlig overvejelse af relationelle aspekter af etik og de sociale dimensioner, der informerer visionen og skabelsen af teknologi. Etik i denne rammesætningbetragtes som den igangværende proces af beslutningstagen baseret på moralske vurderinger, og mit arbejde forsøger at tage denne undersøgelse ud over udforskningen af normative vurderinger og forskriftsmæssige retningslinjer, men ind i felterne af forestillede fremtider og affektive oplevelser af moralske spændinger og værdikonflikter.

Hovedbidraget i denne afhandling er en grundig undersøgelse af, hvad der kræves for, at praktikere bevidst kan engagere sig med etik i praksis, samt en rammesætning for etik i praksis, der inkluderer en konceptuel rammesætning af forestillinger som nødvendige forløbere for etik, og overvejelsen af affektive oplevelser af at træffe moralske beslutninger som nødvendige aspekter at tage i betragtning i et længerevarende perspektiv af kontinuerlig etisk engagement. Denne afhandling giver ikke løsninger på sådanne udfordringer, ej heller endnu en rammesætning til at vejlede praktikere i at træffe bedre beslutninger. I stedet håber jeg at bidrage til en bredere forståelse af, hvad vi skal tage i betragtning, når vi designer til sådanne forventninger, og give vejledninger til, hvordan det kan gøres.

Jeg gør dette ved først at skitsere de produktive påvirkninger af sociotekniske forestillinger på designpraksis, ved at analysere eksisterende casestudier af velmenende designprojekter fra STS-feltet. Jeg fremhæver tre vigtige dimensioner, hvor sådanne nuværende sociotekniske forestillinger forudbestemmer, hvordan teknikere kan en-

gagere sig med værdiorienteret arbejde - initiativer, der sigter mod at vejlede en mere etisk adfærd. Jeg argumenterer for, at hvis underliggende nuværende sociotekniske forestillinger ikke identificeres og klart adresseres, vil alt efterfølgende arbejde med at overholde værdier eller retningslinjer blive tolket til fordel for sådanne eksisterende forestillinger.

For det andet præsenterer jeg et casestudie af en sådan identifikation af nuværende sociotekniske forestillinger i sammenhæng med AI og bæredygtighed. Baseret på dette casestudie præsenterer jeg en rammesætning og et vokabularium, baseret på det filosofiske koncept af "device paradigm" af Borgman, der hjælper med at forstå de forførende dynamikker af sådanne forestillinger og hvordan det påvirker og former forholdet mellem moral og teknologi for teknikere.

For det tredje præsenterer jeg en metode til at engagere sig med og udfordre eksisterende sociotekniske forestillinger sammen med praktikere og forskere inden for sensor-netværk. Denne metode bruger spekulativt rollespil og centrerer værdispændinger gennem omsorgsetikkens linse som en situeret tilgang til at reagere på socialt udfordrende situationer, hvor mennesker møder sårbarheder og de kontekstuelle begrænsninger af teknologi. Denne metode præsenterer en produktiv tilgang til at udfordre forestillinger ved at fremhæve situerethed, performativitet, multiple stemmer og aktører, samt behovets indbyrdes afhængighed.

Endelig engagerer jeg mig i etnografisk arbejde med et team af praktikere over en længere periode, for at undersøge de følelsesmæssige aspekter ved at integrere etiske indsatser i løbende dagligt arbejde. Jeg anvender konceptet; moralsk stress, baseret på tidligere forskning inden for moralpsykologi, ledelsesstudier og sygeplejestudier, til de stadig mere moraliserede arbejdsmæssige opgaver for udviklere, og sporer virkningen af denne oplevelse gennem den daglige arbejdserfaring inden for teknisk praksis. Dette kapitel knytter også tilbage til at påpege, hvor praktikere støder på sociotekniske forestillinger, som eri spil, og den stress det medfører, at ændre tankegange her omkring.

Som sidste bidrag knytter jeg implikationerne af alle studier sammen ved at sige, at for at enhver etisk intervention skal lykkes, skal vi starte med forestillingerne om teknologi og dens indvirkning på socialt og politisk liv, og vi kan ikke stoppe med interventioner for at ændre disse, men vi skal også tage højde for de affektive konsekvenser af sådanne engagementer, som praktikere oplever.

Acknowledgements

As with many long term project, this dissertation is better because of the many people who have supported me throughout the last three years (and a little more). This has been a very collaborative project, so there are many people to mention (though as always, this list is not conclusive):

My supervisor Irina Shklovski, whose unwavering support and believe in my ideas and hunches throughout this project have kept me going. Without your guidance and your dedication I would have never been able to trust this process that led me and this thesis to an unexpected and much better place than I could have imagined when I started it. I feel very lucky that I have had the honour to work with you, and to benefit from your expansive knowledge and sharp questions. I would not have wanted it any other way. I also want to thank my supervisor Marco Rozendaal. There is no sweeter person in DCODE, and I appreciate your care, your open mind and tireless encouragement, that have made me feel seen both as a person and scholar. Together, you both made for the best support team I could ask for, and I am grateful for the engaged conversations we had, about design, ethics, AI and life outside of this PhD.

The DCODE network. Aside from the funding, the network has provided me with valuable context and contacts for my research, and the collaborations and schools across these three years have contributed significantly to my approach to the work you are reading. I want to thank my research colleagues across the many involved institutions, that have shared their expertise and advice with me. In particular my fellow ESR collaborators Youngsil Lee, Rob Collins, Yuxi Liu, Aditi Surana and Carlos Millan, who I had the honor to produce some great research with. I also want to thank my other co-authors that are part of the consortium: Andrea Mauri, Lachlan D Urqhart, John Vines, Cara Wilson and Larissa Pschetz. Also, Mireia and Mugdha - thank you for your friendship beyond DCODE, you made me feel like I belong.

To all the people I met and collaborated with during my projects, at AMS, at ACRC and plenty of other places, I want to extend my gratitude for sharing your work with me, your stories, your concerns and your expertise. To a large extend, this dissertation is for you.

My fellow ethics coven siblings Mirabelle Jones and Natalia Avlona, for the mutual encouragement in difficult times, the many chats and discussions, shared feedback and conversation around our work and beyond, and PhD life in general. Also my thanks to Ville Vakkuri, without whom I would not have found the lens of moral stress to put a name on the thing I have been talking about for two years.

To the HCC community at DIKU, and particular my fellow PhDs and the entire section, a heartfelt Thank You for always being open to help, making shared misery more fun, and the commute across the bridge worthwhile! In no particular order, thank you to Valkyrie, Teresa, Thomas, Tor-Salve, Bhaskar, Sara, Sarah, Kellie, Martin, Barry, Kasper, Johanna, Arpit, Valeria and Athina.

Finally, I want to share some personal thanks. My heartfelt gratitude goes out to my friends, near and far and too many to list, who have listened to me complain, worry and doubt throughout the last three years, just to hear me reiterate in the next sentence that this is really what I want to do and what makes me happy. I probably wouldn't have believed me either, but its true! Thank you for sticking around, and putting up with my weird schedule, deadlines, crunches, and reviews, and believing in me always. I want to especially thank Michelle and Anu, who have inspired me to begin a PhD in the first place and make me believe that I could. You are my PhD fairy god mothers. Thank you to the Block Zero family, for your ongoing support and friendship and interest in the work I do now.

An dieser Stelle möchte ich auch meiner Familie danken, die sich über die letzten drei Jahre meine Monologe anhören mussten. An Mum und Audrey, für eure Geduld (mit meinem zu vollen Kalender), für eure Ermutigungen, Umarmungen und Zuspruch. Insbesondere an Maurice und Ben - eure Schwester zu sein motiviert mich zu mehr als ich mir selbst zutrauen würde.

Lastly, and most importantly, Henning, my nearest and dearest. You kept me sane throughout these years, and without your encouragement, your patience and love I would not be able to think and write as freely as I do. You made this work your world because it was mine. I could write a whole section on what your support means to me, but for now I will limit myself to saying thank you for bringing me snacks and tea in the correct cups.

Publications

Included in this dissertation

Paper 1
Rattay, S., C. Rozendaal, M., and Shklovski, I. (2024) Situating Imaginaries of Ethics in / of / through Design, in Gray, C., Ciliotta Chehade, E., Hekkert, P., Forlano, L., Ciuccarelli, P., Lloyd, P. (eds.), DRS2024: Boston, 23–28 June, Boston, USA. https://doi.org/10.21606/drs.2024.803

Paper 2
Rattay, S., C. Rozendaal, M., and Shklovski, I. "AI as the Final Moral Device"
Submitted December 2023 to Big Data and Society

Paper 3
Sonja Rattay, Robert Collins, Aditi Surana, Youngsil Lee, Yuxi Liu, Andrea Mauri, Lachlan D Urquhart, John Vines, Cara Wilson, Larissa Pschetz, Marco C. Rozendaal, and Irina Shklovski. 2023. Sensing Care Through Design: A Speculative Role-play Approach to "Living with" Sensor-supported Care Networks. In Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23). Association for Computing Machinery, New York, NY, USA, 1660–1675. https://doi.org/10.1145/3563657.3596066

Paper 4
Rattay, S., Vakkuri, V. C. Rozendaal, M., and Shklovski, I. "The Affective Experience of Ethics"
Submitted April 2024 to ACM Nordic Human-Computer Interaction Conference 2024 (NordiCHI)

Contents

1 Introduction

On 16th Jan 2016 I read an article on Roombas that leaked sensitive information about the home they were supposed to keep clean, to advertising services. The article described in chilly detail how the sleek smart home environments dreamed up by designers and developers in RnD departments across big tech companies touch on difficult topics such as privacy, agency and autonomy. The days before, I had just finished an intense two week design sprint with a global home appliance company on developing vision candidates for a smart home strategy. Smart Vacuums, amongst various other sensor equipped appliances, featured in many of the developed user scenarios we presented by the end. Reading the article touched a nerve that day. I generally prided myself as being an ethically aware designer, with strong sensitivities for the ethico-social dimensions of design work (I tried to ensure for example that we would have diverse gender representations in our scenarios taking care of domestic chores) and I was acutely aware and up to date on the rising conversations around the harms of digital technologies. So seeing my recent design activities mirrored in this critical article re-enforced a growing feeling I had carried for a while: that despite my good intentions, *I was* part of a bigger problem.

This reckoning was in line with a major cultural shift within the tech and innovation scene globally, coined the *techlash* in following years (Foroohar, 2018), fueled by media reporting on concerns over issues such as privacy breaches (Schechner, 2019; Braithwaite, 2018), user manipulation (McEvers, 2017; Lewis, 2017), the spread of misinformation (Doubek, 2017; Rosenberg, 2018), and the monopolistic power of tech giants like Facebook, Google, and Amazon (Bond, 2020), along with multiple consequential publications which pushed such issues into mainstream awareness (O'Neil, 2016; Chang, 2019; Wachter-Boettcher, 2017). The increased scrutiny created a tipping point in mass awareness around the critique of the early 2000s Silicon Valley culture and the frustrations with large tech companies (Gardels, 2018). Part of the critiques was a deeper examination of mainstream industry design practices, predominantly shaped by Design Thinking. Design Thinking, as pioneered by IDEO and d.school, is supposed to enable innovation within the field of digital technologies (Corral and Fronza, 2018; Culén and Følstad, 2014), by merging desirable outcomes with technical feasibility and business interests through an iterative process of researching, ideating, testing and refining (IDEO, 2022). While Design Thinking is claimed to be value-neutral, and making space for the consideration of ethical concerns as part of the research of user needs and testing of systems, it has been under plenty of criticism for contributing to the frustrations with tech development (Irani, 2018), aligning itself with the general mantra of "moving fast and breaking things" (Taplin, 2017). The modularisation and commodification of the process driven by large design consultancies such as Frog, Fjord and the original driver of the business oriented flavor of design thinking IDEO, has led to an absorption of fragments of the original idea into various cracks of business practice, where it seems to fit with the existing processes and structures (, 2021). Rather than creating a larger shift and challenging business infrastructures incompatible with deeper design maturity, design thinking has been watered down into easily digestible bits and pieces, focusing on selling design as good for business, and targeting the various tools of design towards selling points such as user retention, ongoing user engagement, and constant growth of products, rather than for lived experiences (Ackermann, 2023).

This shift of perception also affected the people working for these companies and related areas, and designers and developers started to take more note of the ways in which their work contributed to the controversies (Su *et al.*, 2021). While I was never part of any of the companies under fire during the *techlash* (in fact, I could only dream of working for them one day), I followed the conversations on design blogs, forums and events eagerly. In 2015 I had co-founded a design and tech studio, that grew into a specialisation for strategic proof of concept projects, often as parts of what our clients understood as their innovation efforts[1]. We develop design strategies, research market conditions and build prototypes and demonstrations for next generation products for local start ups as well as global tech companies. In particular we focused on expert systems, not consumer products - advanced and complicated applications to be used in critical environments, such as energy, finance, health care or global logistics. Across these fields, I tried to find the touch points between my work as head of design and the controversies covered by prominent designers influencing my understanding of the industry.

Ethics had become a prominent focus object more than ever, with increasing attention on how to design technologies responsibly (Zhang, 2023). Design ethics have become a theme in itself, with events, networks (CTH, 2024; EDN, 2024) and resources (Gispen, 2017; DDC, 2021) aiming to raise awareness around the ethical dimensions of the design of technology and provide guidance for how to attend to them. Prominent voices in the community published opinionated books (Monteiro, 2019) and ted talks (CTH, 2024) about the political responsibilities of designers. I was also inspired by Design Justice (Costanza-Chock, 2018), and its core principles focusing on community, equitable participation, marginalized voices and accountability to the affected. However, I also quickly realized that it was impossible for projects like the ones we were working on to involve directly impacted communities, even though our strategies might very well lay the ground for product decisions down the line.

[1] Proof of concept projects provide research, demonstrations and prototypes to develop strategies for technologies that companies aim to bring to market in the next few years or decade. These projects don't aim for immediate market release but to inform on what it would take for developing what might become the next generation of products, and a longer term strategy to work towards.

Eventually, I had to concede that trying to make the world a better place while running a service-oriented business is a big challenge. My co-founders and my team where as open-minded and invested as I could ask for, and our clients trusted us and gave us plenty of leeway to approach our collaborations however we saw fit - but in the end, some market conditions and business requirements demand trade offs and concessions that made me realise how idealistic and naive certain notions about a responsible design process are[2], and how difficult it is to draw the lines between politically motivated design frameworks and the design of complex expert systems. So when I got the opportunity to go full research into how to approach design ethics in tech development practice, I was delighted.

My unofficial research question that motivated me to enter this PhD program was how to be a designer that does not screw over the world by designing technologies that are superficially satisfying but ultimately harmful, or by following flashy market trends in terms of practice and design fields without really understanding how our methods and practices actually affect the people we design for and with beyond the frames of workshops, focus groups and user interviews. Coming from those 5 years of co-running a design studio and being influenced by business and marketing thinking, investigating this concern from an academic perspective felt daunting, but also like the only reasonable thing to do - Understanding the dynamics of industrial design practice while being deeply embedded, a driver even as prominent speaker and thinker in the local design community, was deeply frustrating, confusing and to be frank - personally destructive. At the same time, I figured three years to focus on this concern would be plenty, coming from a field in which projects are scheduled with rapid progress, and time is always tight. Three years to be left alone from the demands of running a business, not being responsible for paying salaries and managing clients would give me the freedom of thought to find the missing parts that would fix the problems I had encountered of putting "ethics into design practice".

I admit that I was somewhat naive.

[2] One example is the common demand to involve many stakeholders and affected. While we were certainly aware that the systems we worked with would affect the general population, working with 'all stakeholders affected' is a logistically unrealistic approach, when most designers have trouble getting the lead developers within an organisation at the same table across locations and time zones. Many of our projects targeted professional expert users, so finding contact points with directly affected was a big challenge. When designing a city wide district heating system for example, few citizens will have a direct point of interaction with the system, even though all will be directly affected.

1.1 Challenges of Designing the Digital

Designing digital products and services has always come with challenges of various kinds. The design of technologies is always an endeavour of creating better futures, by introducing change that bring us closer to an idea of a good life (Löwgren and Stolterman, 2007). This orientation towards the future creates both thrill about new opportunities and concerns about unintended consequences. In this sense, designing technology has always been about ideas of a better life, in one way or another. The notion of what constitutes such a better life with technology however is a fuzzy one.

The development of considerations in HCI is often characterised in waves, with each wave shifting the concern of what is necessary to include when addressing interaction between humans and technology. Bødker (2015) describes the transition from second-wave HCI, which centers on usability in work-related contexts, to third-wave HCI, which encompasses broader, more diverse contexts emphasizing experience, meaning, and socio-cultural aspects. Höök and Löwgren (2021) highlight that the digital as design material in focus itself provides not a very stable focus point, given its rapid evolution both in function, capabilities and entanglement with social structures. Especially those entanglements are reason to consider ethical considerations beyond traditional human-centred approaches (Frauenberger, 2020). Bannon (2011) emphasizes the need for HCI to prioritize human needs, interactions, and social contexts in the design of technology. He argues that instead of forcing humans to adapt to technological constraints, HCI should focus on creating technology that seamlessly integrates into and enhances daily life, ultimately fostering a more symbiotic relationship between people and technology.

Launched in 1986, human-centred design (IDF, 2024c) most prominently articulated the principle that user-needs should be a core concern when designing digital technologies, to account for the fact that digital systems were increasingly integrated in everyday life situations and adopted by non-expert users. From usability to pleasurable user experiences, from intuitive interfaces to privacy concerns (Bødker, 2015), what has been considered *'good'* design of such systems has been a shifting target for years. In comparison, Universal Design (IDF, 2024b) recognises that humans differ in their needs, capabilities and skills, and that people with needs that diverge from the ergonomic norm should be prioritised, to make systems accessible to everyone. Circular Design (IDF, 2024a; EMAF, 2024) brings questions of resource management into the equation, and aims to also address ecological conditions of design. These two are examples of design strategies that bring concrete priorities to the table when considering positive contributions of design beyond general usefulness. Participatory Design (Co-design) and Value Sensitive Design on the other hand put their focus more on ways of designing, and the how of inclusion of different stakeholders. Participatory design, in its original Scandinavian tradition (Bjerknes *et al.*, 1987), moved political dimensions of work place technologies into focus (the less politically charged version has been named co-design, to capture the pro-active involvement of users beyond initial user research), while Value Sensitive Design developed processes that aims to explicitly name and intentionally attend to potential values considered in the design of technologies - such as privacy, transparency or participation (Friedman and Kahn, 2007).

1.1.1 Designing AI

In addition to the broadening discourse around ethics in design, accelerated technical progress has added to the complexity of the discussion - in particular the current hype around the rapid advancements of AI technologies[3]. AI therefore is becoming an increasingly discussed topic in HCI, both for the hope that AI can provide opportunities for new and improved user experiences (Dove *et al.*, 2017; Holmquist, 2017; Grudin, 2009), as well as for worries and concerns around harms it might facilitate, and the responsibilities designers take on for that. While HCI and design routinely deals with complicated systems and advanced technologies as laid out above, it seems that AI brings challenges for established design tools and processes such as sketching, prototyping and testing, which translates to designers struggling with engaging with AI as a design material, especially in regards to envisioning and prototyping AI systems (Gillies *et al.*, 2016; Holmquist, 2017; Allen, 2018). Yang and colleagues (2020) argue that this is due to two distinctive design challenges: The uncertainty around AI capabilities and the adaptive complexity of output. These factors present challenges to the tools designers typically apply to investigate potential consequences and mitigate undesired outcomes.

The unpacking of potential socio-political harms of the spreading use of AI has led to launches of research initiatives, academic conferences, governance efforts and industrial efforts, which can be broadly pulled together under the umbrella of *AI ethics*. The field of AI ethics includes diverse actors in a variety of sectors (Ortega-Bolaños *et al.*, 2024). Work in this field generally covers work on fairness (Mehrabi *et al.*, 2022), accountability (Raji *et al.*, 2020; Novelli *et al.*, 2023), transparency and explainability (Calvi and Kotzinos, 2023), bias (Pereira, 2020), and justice (Vakil, 2018) within algorithmic systems, with the hopes of reducing harms, biases and discrimination caused through such systems. Examples are technology companies that develop their own guidelines and toolkits (Young *et al.*, 2020; Pair, 2024), professional and academic organisations that aim to develop shared codes of conduct across the field (Shahriari and Shahriari, 2017; Whittaker, 2018; Internet Society, 2017), governance and legislators working on policies and regulations for AI technologies (EU, 2021; EU, 2019; OECD, 2024; CA, 2023), researchers across a multitude of disciplines - technical, legal, philosophical etc - producing work that informs the prior efforts

[3] The definition of "AI" is a continuously contested one Dobrev, 2012. The technical boundary of AI, even in AI research communities, is disputed and continuously evolving (Stone *et al.*, 2022). For the sake of this thesis, I will adopt the definition of Kaplan and colleagues (2019): "AI refers to computational systems that interpret external data, learn from such data, and use those learnings to achieve specific goals and tasks through flexible adaptation"

(Morley *et al.*, 2023; Brundage *et al.*, 2018), and lastly consultancies and independent designers, creating their own tools in the forms of methods or services (PWC, 2024; IDEO, 2019; Tangible, 2024). The field therefore contains a multitude of voices, interests and motivations, that sometimes align, sometimes come into open conflict, and sometimes appear to align on the surface but turn out to contradict each other when being investigated more thoroughly on their sociopolitical implications.

1.2 Contribution

Designing ethically, ethical AI and where both meet - designing ethical AI. Academic efforts in this area are limited, and encounter the same issues as academic design ethics and AI ethics efforts - a disconnect between ideal notions of design and development practice, and a disregard of the gap between high level ethical concerns and applied practice. Industry efforts encounter the same issues as industry efforts in regards to ethical design and AI ethics - a confusion about vocabulary and conformance to market practices which leads to superficial engagement with ethical dilemmas. In this field of tensions, I start my investigations at the intersection of Design, AI and Philosophy, to find a path forward in answer to my overarching research question of how to deal with ethics in practice for the design of AI systems. I break this question down in four sub questions, which I will elaborate on below.

1.2.1 Where to start

Ethics and design are inherently connected, as the former is about decision making in regards to more desirable futures, and the latter the activity of making such futures happen (Devon and Poel, 2004). Understanding why ethical tools do not have the desired impact requires investigation into the assumptions they are build on. Ethics, or moral philosophy, as it turns out is hardly a straight forward endeavour. There are many rivaling theories of ethics, each bringing forth a different conception of how we ought to live in order to achieve a "good life", a moral life. Since Socrates presented his definition of morality in 390 BC, nobody can agree on which theory is ’the one’ to answer the questions of moral philosophy after what makes up a ’good life’. Nevertheless, ethics have become the focus point around which people gather to ensure that certain promises of technology are fulfilled, while worrying about minimising harms. In this sense, technologists worry about an unfocused notion of ethics, as a concept that somehow points us towards something being done potentially wrong or not exactly right, in particular when things - mostly referred to as unintended consequences - turn out harmful. This concern around the ethical impact of information technology in particular have been debated since Moor raised the field of computer ethics (Moor, 1985), and traces through various ups and downs along different hype cycles - the latest one being AI ethics - until today, in similar circles.

I therefore start my research in a different place. Tech ethics begin way before designing technologies, they begin when we imagine what our future with technologies might look like, and which promises technologies will fulfill in them. Design, as a world making activity that creates intentional changes with specific outcomes in mind, fundamentally relies on imagination and envisioned futures. STS scholars have captured the formative forces of imagined futures in the concept of imaginaries (for example Jasanoff, 2015; Ames, 2019; Dahlgren *et al.*, 2021; Sadowski and Bendor, 2019; Miller, 2020; Lustig, 2019 to name a few). Social Imaginaries, as conceptualized by Taylor (2002), are interpretations of reality and potential futures shared by many people. They describe the mental frameworks communities use to understand and make sense of the world together. They provide the narrative infrastructure within which people contextualize their experiences and explain the environments and systems they inhabit, both individually and in relation to each other. These shared understandings and interpretations affect how humans approach challenges and issues in the world we live in and from which perspective we legitimize specific solutions for them.

In this sense, imaginaries provide a bridge between abstract moral conceptions and desirable futures we can imagine at present. These underlying imaginaries of technology establish mental boundaries for ethical decision making (Markham, 2021). By providing aspirational and normative frames in the shared perception of futures, imaginaries contain collaboratively accepted visions of how life ought or not ought to be lived. By shaping and limiting the ways in which designers make sense of their socio-technical environments, imaginaries become mundane frames of world-making, which manifest themselves in the design motivations and design decisions designers make in their work. In other words, they express a shared understanding of what design actions are legitimate to reach desirable futures, thereby encoding moral deliberation within its visions of those futures. The effects of such encoded normative qualities have been well documented by STS scholars who have mapped formative socio-technical imaginaries surrounding data driven technologies. These imaginaries provide a useful framing for the ways in which visions of a better future are intricately connected to visions of technology, and thereby provide grounding of possibilities from which conditions for ethical decision making grow. Given their structuring forces on futures, socio-technical imaginaries impact the design of technologies by shaping what is considered technically desirable and feasible. In turn, technology design becomes an expression of such imagined futures, and materialises what technologists consider aspirational and unquestionable.

I therefore use the concept of socio-technical imaginaries as a lens to investigate the influence imagined futures have on tech ethics. I start with the hypothesis that in order to affect ethics in practice, we need to understand how designers and developers understand and relate to ethics beyond their practice. I therefore start with two research questions:

1. *How does the way practitioners think about ethics enable or preclude ethical action?*
2. *How do designers imagine AI to factor into ethics?*

1.2.2 Ethics IRL

There has been much work on frameworks, guidelines, codes of conduct and toolkits to provide guidance and inspiration for ethical design of tech in practice, or what I call Ethics IRL[4]. They reach from collections of high-level principles to processes and methods for more concrete ethical considerations. There are technical tools that aim to provide measurable metrics for algorithms to measure performance on quantified interpretations of ethical issues, such as bias or fairness markers, and tools that are anchored in design practice and providing modified design tools for ethical ideation and consequence scanning. There is Value Sensitive Design (which I will shortly elaborate on in chapter 8), which despite its long [5] history struggles with broader uptake, and similar process oriented interventions, that seem to try to re-invent the wheel. It appears that amongst this variety of proposed interventions, we still get stuck. In fact, it rather seems that the pressure to find creative ways of engaging ethics is increasing, with the escalating complexity of digital networks and technological ecosystems, encompassing various human, non-human and artificial actors. So if we manage to make the bridge between value work in design and the imaginaries that drive tech and AI development, we also need to tie such understanding into formats of engagement that push beyond reflection and solutionist thinking, which leads me to my third research question, following the prior two:

3. *How can we design formats of engagement for practitioners to work through the ethics of imagined tech futures?*

However, thinking back to my original, personal struggles of connecting my ethical concerns to the reality of my everyday work context, developing such formats is still not enough. Frameworks and guidelines such as the ones described previously struggle with contextual interpretations of ethical decision making, when practical everyday life comes in between good intentions and applied implementation. What is considered in line with these ethical frameworks differs between service

[4] IRL is the abbreviation for In Real Life, often used in online and gaming communications, denoting life that happens offline, in the physical space, "for real". I use the abbreviation to indicate the contrast between ideals of practicing ethics and real implementations of ethics efforts

[5] comparably to the other examples which are mostly developed in the last 5 years

providers, producers, clients and customers. As each of them have their own frame of expectations and interpretations, opposing motivations create a tug of war about responsibility and accountability within which design practitioners are caught. As service providers, they are responsible to act in the best interest of their clients and customers. As creators, they face the question as to how far they are morally responsible for the consequences of their creations and their effect on others. These limitations and political ecosystems are as much part of the reality designers need to navigate when being confronted with ethical design decisions, as design processes and tools. The core of their dilemma consists in exploring and identifying the interactions between conflicting requirements and expectations and deciding on an acceptable compromise (Carstensen and Schmidt, 2016) on behalf of the involved actors. More dynamic, engaged modes of reflection might create spaces for negotiating such compromises in less stringent manners, but plenty of work remains to bring such compromises to fruition. The last research question of my thesis is therefore:

> *4. How do practitioners experience the doing of ethics in an organizational everyday context, and how does that affect ethical practice?*

This brings us back to the original story, one of a designer trying to do “the right thing”. We end back with what people do in the mundane every day, because that is where the conceptual notions of ethics tend to fall apart – the specific, concrete, partial, subjective and unique context of each individual design decision, from tiny (the size of buttons) to massive (which data do we use to classify someone a criminal or not). But these unique situations can not give us the insight created from oversight of many scenarios – and so we need both. The exploration of future potentials that lead us to high-level, principled guides towards something, and the recognition of the applied, intricate and ultimately human facets that unfold when pulling the high-level to the office floor.

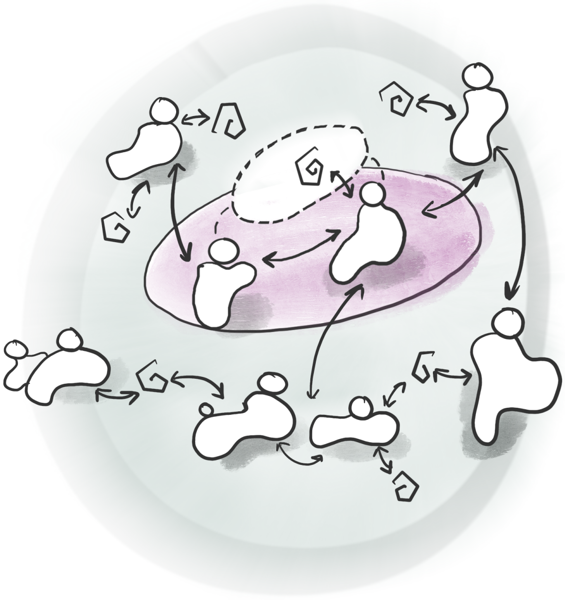


Figure 1.1: Illustration of the relational complexity of navigating Ethics in Practice, from my notes

2 Reader's Guide

This thesis is structured as followed:

Part 0 includes the introduction, this readers guide and a description of the research context for my project, as well as a statement of positionality, that is relevant for framing what prior experience I bring to this research.

Part 1 and part 2 each contain 4 chapters - one background chapter that summarises the relevant related work for this part, two chapters based on prior research papers, and a conclusion.

Part 1 addresses the following two research questions:
1. How does the way practitioners think about ethics enable or preclude ethical action?
2. How do designers imagine AI to factor into ethics?

Part 2 addresses the other two research questions:
3. How can we design formats of engagement for practitioners to work through the ethics of imagined tech futures?
4. How do practitioners experience the doing of ethics in an organizational everyday context, and how does that affect ethical practice?

Part 3 includes the overall discussion and conclusion for this dissertation.
Part 4 includes the manuscripts of the papers covered in this thesis.

This structure is somewhat unusual for a dissertation within HCI. For one, it does not have a dedicated methods chapter. Instead, the methods used for each study are covered in the individual paper-based chapters, as part of the summary of the study, and the implications for this thesis. I made this choice because the methods used for each project are very research context depended, and vary widely. Presenting all methods used in each study in one chapter would have resulted in a chapter with four distinct sections to explain each method and its relevance and appropriateness in relation to the study context, and hence would have ended repeating a lot of information covered again in the paper chapters, and also be more confusing to follow. In addition, the methods are closely tied to the implications presented from each paper. It therefore makes sense to present methods, insights and implications according to study context, in favor of the flow of the argument, rather than by content type.

The main insights produced during this PhD project are covered in part 1 and part 2. This split of chapters that cover the produced work is also unusual, but makes sense for the structure of the arguments I am making throughout this thesis - both paper chapters in part 1 contribute to my imaginary-related work, while both paper chapters in part 2 address practice concerns. As such, the parts build on each other, and require different scaffolding and framing through related work and conclusions, which lead to the next parts.

3 Research Context and Positionality

3.1 DCODE

This project is part of the European research network DCODE, a Horizon 2020 Marie Sklodowska-Curie Training Network of seven European universities, bringing together over 40 researchers over a project timeline of 4 years. The network provides 15 PhD projects, throughout which the consortium collaborates with partners from industry, government and civil society. The mission of this network is to respond to the digital transformation of society by fundamentally re-thinking design in its many layers in five work packages:

1. Inclusive Digital Futures - how will we design for human-machine relations?

2. Trusted Interactions - How will we make decentralised systems work for society?

3. Sustainable socio-economical models - how will we co-create sustainable business models in a digital society

4. Democratic Data Governance - How will we enable public deliberation on data and algorithms

5. Future Design practices - How hill we prototype responsible design practices in the digital society?

Across these 5 challenges (see fig 3.1), 15 PhD researchers develop research projects which aim to proactively steer our future towards sustainability and inclusivity. The DCODE project strives for real-world impact, and aside from a network of host universities, DCODE has a number of non-academic partners, which are open to research collaborations with the PhD researchers. In particular the set up of prototeams created an international and interdisciplinary framing for research collaborations. Prototeams are teams of PhD researchers from different work packages, who collaborate with non-academic research partners to develop scientific knowledge and prototype professional practices. In addition, the network organises twice-yearly

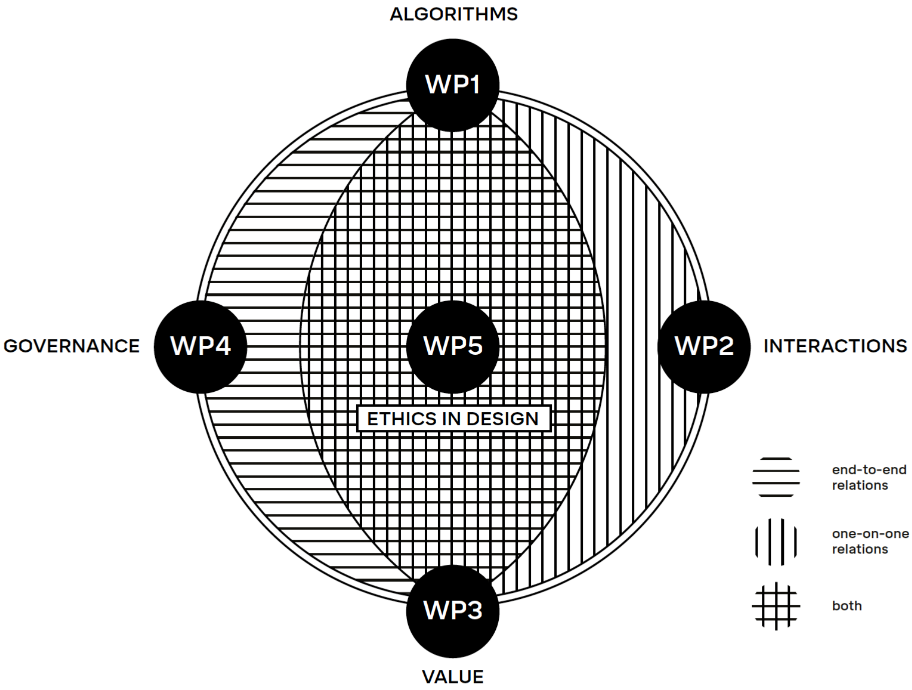


Figure 3.1: Visual of the 5 workpackages within the DCODE research network

PHD schools in which PhD researchers and partners connect about the ongoing research, and work through relevant questions and concepts together.

3.1.1 Project ESR14

My particular project falls under the work package of future design practices, and contributes to the investigation of channeling ethical insights into future practices. In particular, this project explores how ethical standpoints are embedded in design practices for autonomous systems. My project responds to the question of future design practices by mapping out the different layers of ethical engagement in design.

The DCODE network has given me access to research opportunities and collaborations across Europe, that would have otherwise been difficult to access. In particular the work of my first prototeam has contributed to my research, and the project has resulted in the publication that informs my contribution in chapter 9, Designing for Care. Because of the collaborative set up, and the existing connections with the nonacademic partners, it was possible for this project to quickly iterate through different formats of the workshop, and collaborate tightly with institutions that provided a rich research context for the team to explore the overlappings between care and design. The existing network also provided me with the contacts to join the innovation team of a municipality working on smart city technologies and to conduct ethnography on ethics toolkits in practice, which resulted in the study covered in chapter 10. My strong focus on practice is therefore enabled because of DCODEs commitment to real world impact, and the opportunities DCODE has provided to engage with experts beyond academia.

3.2 Positionality

My interest in the theory-practice gap is motivated by my own industry experience as designer, and co-founder of a design studio. Through this professional experience in a full-time leadership position I can relate to and evaluate stories shared by the the people I collaborate with.

At this point in time, designers and developers in the field of innovation that I am considering often work in tight collaboration across the development process. Developers are part of ideation workshops, review user research with designers to understand design decisions, and designers collaborate with developers to circumvent technical barriers to product ideas. For my intends and purposes, their field of practice is closely connected, and similarly impactful to the products in question. As such I consider them one group. Of course each group comes with their very own skillset, and is shaped by strongly divergent education and backgrounds. But the important practice of interest for this thesis is where these skills and mindsets and worldviews come together to enable the production of digital systems. And this coming together happens in the same sphere, shaped by similar industry narratives, organisational structures, imaginaries and logistical concerns. I also want to contradict the notion that designers are more responsible in their conduct towards users, or developers wholly disinterested in the social meaning of their work. Throughout both my professional career and my research, I have met designers so fully embedded in the silicon valley mindset of moving fast and breaking things, and obsessed with glossy, addictive interfaces without paying any thought to the user needs beyond how to wrap them up in their product. And I have met developers that are deeply concerned about their contribution to society and their role in the systems they help bring to market. Considering design practice and hence designers inherently further progressed in some kind of moral route is therefore both naive and offensive. In fact, from my own experience, it is sometimes easier for designers than for developers to hide behind the notion of user centered design and the fact that they do limited user research, to justify themselves, that what they do is inherently good.

My experiences in developing and leading design teams, managing client accounts and collaborating with developers across the entire project cycle gives me concrete contexts to connect to the stories shared by my research collaborators. It also gives me sensitivity to the underlying tensions and challenges that are being expressed in particular vocabularies. My familiarity with this vocabulary and project structures allows me to describe in detail what I observe in the field. It also gives me opportunities and possibilities to collaborate seamlessly and effortlessly in the contexts that I research. At the same time, this familiarity is a barrier to recognise particular strangenesses in the field, that I might already have internalised from prior working habits. I draw from other researchers that have researched the profession of design to find new language to pinpoint and contrast my observations, such as imaginaries to frame the power of envisionings and futures in design, and scholars of ethics in practice to highlight how theoretical ethics work connects and disconnects from how designers *do ethics in the field.*

Part I

Encountering Ethics through Imaginaries

4 Theories of Ethics in HCI

While the term "ethics" and "ethical" seem currently ubiquitous in the discourse around AI futures [1], which ethics in philosophical terms are being referred to is often omitted from such statements (Bietti, 2020; Lauer, 2021). That doesn't mean that there aren't ethical frames of thinking at work, there are just not denoted intentionally outside of discourses driven by philosophers (Steen, 2015). As Shilton (2018) points out, considerations around values and ethics in design quickly encounter challenges of vocabulary and terminology, as such work spans disciplines across design and philosophy (Ozkaramanli *et al.*, 2024). For this thesis I follow Shilton (2018, p.21) in her definition of Values and Ethics: Values as goals that are desirable, worthwhile or good, and are perceived to be broadly applicable to social life, and Ethics as the discussion of ethical frameworks, which aim to guide a decision making process to identify morally right and wrong actions, partly guided by present values. So in order to pin down what the term "ethics" in the related work might stand for, I will start with a brief mapping of the potential candidates of ethical frameworks and approaches that have informed ethical engagement in HCI.

4.1 Consequentialism, Deontology, Virtue

Generally, normative ethical theories in the western tradition are broken down into three strains: deontology, consequentialism, and virtue ethics. Many approaches dealing with ethics relevant to HCI fall within these three categories, starting with Moor's (1985) influential monograph on computer ethics and Johnson's textbook (2000) on related controversies. Moor's paper "What is Computer Ethics?" highlighted the issue of policy vacuums, where new capabilities of computers lack ethical or legal guidelines. Moors' concept of logical malleability (Moor, 1985) highlights the versatility of computers and their ability to be used in a vast array of applications, raising, he contends, unique ethical challenges due to their flexibility. In addition, he argues that the rapid advancement of computer technology outpaces the development of appropriate policies and ethical guidelines, creating a "policy vacuum" that needs to be filled with new ethical frameworks. To meet this vacuum, he suggests Just

[1] Along with the terms responsible, fair, secure, safe etc. it often remains unspecified which ethical foundations underlie the efforts designers and developers are expected to take to create "ethical" technologies, and ethical AI in particular. "Ethical" seems to often be used as a sub category of the listed words, denoting the assumption that "ethical" is just one of the characteristics tech development needs to take on in order to be responsible and safe.

Consequentialism (Moor, 1999), a hybrid framework that merges the outcome-focused approach of consequentialism with the principle-based considerations of deontological ethics. It aims to enable the maximisation of greater good for society by reviewing the outcomes of computer technology, while also taking into account principles of justice and fairness and individual rights to privacy, autonomy and freedom. While just consequentialism aims for utilitaristic outcomes, they must be achieved through just means. Building on Moor and Johnson's work, computer ethics has evolved. For example, Brey (2000) suggests disclosive computer ethics as an anticipatory approach that focuses on uncovering embedded values in technologies during their design, and encourages the selection of appropriate values for a given technology and to evaluating the adherence with the suggested principles throughout the design process (Brey, 2012). Reviews of ongoing research within machine ethics on the use of ethical frameworks report that consequentialism and deontology remain the most prevalent frameworks for ethical consideration in machine ethics (Zoshak and Dew, 2021; Tolmeijer *et al.*, 2021).

However, either approach encounters challenges when operationalised in practice. Consequentialist approaches face the problem of identifying relevant consequences, and finding an appropriate level of detail and scope to consider (Card and Smith, 2020). In real-life situations, it is impossible to identify and address all possible consequences, especially since causations of consequences are not necessarily clear cut - multiple things can contribute to a specific consequence to various extents, just as one thing can have multiple consequences. Even if consequences could be accurately and comprehensively predicted, they would need to be properly quantified in terms of their desirability, with generalisable strategies for comparison or aggregation of the impact of various consequences. However, when we talk about socio-political impact of technologies, such quantification and calculations of greater good presents various ontological and political challenges (Cunningham *et al.*, 2023). Operationalising a deontologist approach also encounters multiple issues. Aside from deciding which rules to implement, this approach assumes that all necessary rules will be possible to identify before being applied. Finding the right level of detail is also challenging. Since rules are expected to be strictly followed, each exception requires an amendment, potentially leading to excessively long and complex rules. However, if the rules lack the appropriate level of detail in their description, or the opposite, are too specifically formulated, they become uninterpretable for a machine (Bonnemains *et al.*, 2018). Additionally, there can be conflicts between rules, either in general or in specific situations. Although ordering or prioritizing rules can address this issue from an implementation standpoint, determining their relative importance is challenging.

Vallor (2016) therefore argues that deontology and utilitarianism are insufficient for dealing with the complexities and rapid advancements in technology. Instead, she suggests that virtue ethics, which emphasizes the development of moral character, offers a more flexible and human-centered alternative. Virtues are about habitual practice of characteristics that guide our actions and decisions, shaping our moral character over time, to achieve moral excellence and a well-lived life. What sets virtue ethics apart from consequentialism and deontology is its emphasis on virtue as the core element of the theory (Kawall, 2008). Consequentialists define virtues as traits that lead to good outcomes, while deontologists see them as traits of those who consistently fulfill their duties. In contrast, virtue ethics resist defining virtues through more fundamental concepts. Instead, virtues and vices form the foundation of virtue ethical theories, with other normative notions being derived from them. By fostering moral virtues, Vallor argues, individuals and societies can better navigate the complexities of the digital age and create a future worth wanting. Virtue ethics have been used to investigate (Ustek-Spilda *et al.*, 2019) and inform (Hagendorff, 2020; Hagendorff, 2022; Bilal *et al.*, 2020) individual conduct for and by technical professionals. Similar to consequentialism and deontology, virtue ethics encounters the issue to define which virtues should inform the design of technology. In addition, judging whether an act is virtuous cannot be done merely by observing actions; the underlying reasons must be clear (Sreenivasan, 2002). If applied to technology, measuring how specific virtues affect the design and the consequential impact of technologies is a challenge. Even if designers where to be perfectly virtuous, resulting technologies and systemic effects might still create harmful outcomes.

4.2 Values and Principles

Value sets and principles are mostly dominant in areas of technical practice, rather than fields of design (Khan *et al.*, 2022; Fjeld *et al.*, 2020; Floridi *et al.*, 2018). They provide prescriptive modes of ethics formulations, which aim to provide a code of conduct for practitioners to comply with. For example, principle-based frameworks such as by van de Poel (2016) and Brey (2012) encourage the selection of appropriate values for a given technology and present example cases to demonstrate the importance of evaluating the adherence with the suggested principles throughout the design process.

A framework for value inquiry connected to design of technologies that has been quite successfully recognised, at least is academic and governance spaces, is Value Sensitive Design (VSD) (Friedman and Kahn, 2007). VSD connects value discovery with tools and methods to bring value alignment into the design process, aiming to foreground human wellbeing and advance human flourishing (Friedman and Hendry, 2019). VSD has furthered the conversations around the role of values in the design of technology and has been used in several different contexts (Friedman *et al.*, 2017), and more recently also in the development of AI (Umbrello and Poel, 2021).

Feminist, anti-racist, and postcolonial scholarship critiques how prior listed approaches often lead to technologies which represent values of dominant social groups, while neglecting those of marginalized communities. Bardzell's work on feminist HCI for example argues for the incorporation of feminist values like "agency, fulfillment, identity and the self, equity, empowerment, diversity, and social justice"

(2010, p.1301). HCI scholars also examine the influence of intersectional identities on sociotechnical experiences, such as class (Ames *et al.*, 2011), race (Hankerson *et al.*, 2016; Schlesinger *et al.*, 2021), gender (Fiesler *et al.*, 2016; Light, 2011) and neurodivergence (Waycott *et al.*, 2015), while postcolonial scholars criticize Euro-American-centric perspectives in HCI (Bidwell, 2016) and argue that postcolonial theory can help us understand the importance of intercultural values in design (Srinivasan, 2017), highlighting historical power relations and cultural hybridity (Irani and Dourish, 2009; Irani *et al.*, 2010; Mainsah and Morrison, 2014).

4.3 Ethics of Care

Care ethics have originated in the feminist critique of traditional ethical theories in the late 20th century (Larrabee, 1993. Pioneering work by scholars such as Gilligan (1982) played a crucial role in its development, highlighting differences in moral reasoning between men and women, with women tending to emphasize care and relationships over abstract principles of justice. Further influences are the works of Noddings (2013), who articulated the central role of caring relationships in moral life, and Tronto (2010), who expanded the concept to include political and social dimensions. These scholars argued for an ethical approach that recognizes the importance of empathy, compassion, and the interdependence of human beings.

Care ethics thus challenges the traditional emphasis on autonomy and individualism of prior frameworks, proposing instead that moral understanding and behavior arise from the context of caring and nurturing relationships. Unlike traditional ethical theories that focus on abstract principles or rules, care ethics centers on the concrete experiences and needs of individuals, particularly those who are vulnerable or dependent (Ruckenstein and Turunen, 2019). It advocates for the recognition and appreciation of the emotional and relational dimensions of ethical decision-making, highlighting how empathy, responsibility, and responsiveness to others' needs are crucial components of moral action. This approach challenges the idea of detached impartiality, proposing instead that moral understanding and ethical behaviors arise from the context of caring relationships, where attention to specific situations and the well-being of others is paramount. Recently, care ethics (and other relational approaches to ethics) have also been discussed in the field of AI ethics, arguing that it is relational and requires a social orientation that moves away from western traditional philosophical approaches (Dignum, 2022; Birhane, 2021). I will write more about care in HCI in chapter 9.

4.4 Conclusion

The long ongoing conversation around ethics in technology highlights the continuously developing challenges technical practice encounters. As Shilton states in her extensive review of values and ethics in HCI, "if there were clear rules to follow, HCI would have long ago demonstrated how to avoid biased design" (2018, p.3). However, even in the same context, different moral frameworks can lead to different design results, and any ethical dilemma comes with a trade off, otherwise it wouldn't be a dilemma. And while principles provide a good starting point for ethical discussion and consideration, multiple scholars have pointed out that a focus on high level ethical frameworks risks covering up political disagreement rather than enabling alignment, and that translating them into practice is a non-trivial task (Mittelstadt, 2019) that will inevitable bring up tensions and conflicts (Whittlestone *et al.*, 2019) around underlying motivations and interests. As Simon et al. (2020) state, this requires a broader review of the socio-technical systems in which technologies are situated, not only in academia, but also, just as importantly, within industry practice.

In the next chapter I present socio-technical imaginaries as a way to provide such a broader view. Socio-technical imaginaries are collectively held visions of desirable futures that incorporate both social and technological elements (Jasanoff, 2015). They are frameworks through which societies understand the possibilities, limitations, and ethical considerations of technological advancements. They also contribute to what is considered a desirable and feasible future, and thereby influence which potentials are considered for ethical decision making. I argue that different imaginaries also contribute to the way in which moral frameworks are being applied, or which values are being considered. Imaginaries provide storylines that provide justifications as to why certain values are considered relevant and desirable and paint the futures available, when certain values are being pursued. They therefore open up the conversation around ethical decision making from how (guided by ethical frameworks) to what is being considered in the first place.

5 Imaginaries of Values in / of / through Design

How do we address socio-political imagination and what effect does it actually have on ethics in design of AI systems? This question drove my initial investigation into imaginaries embedded in design practice, which resulted in paper 1, "Imaginaries of Values in / of / through Design", presented at DRS in Boston in 2024. The main argument of the paper centers on the idea that imaginations of technology and its integration into and benefit for society pre-define and direct the way in which designers engage with processes, tools and methods meant for ethical engagement. Digital technology tends to be charged with a certain charisma, in particular as it exists in ideas and imaginaries about the future (Ames, 2015). This leads me to ask:

How does the way practitioners think about ethics enable or preclude ethical action?

5.1 Socio-technical Imaginaries

For this purpose I propose to engage with the imaginaries that are motivating design intentions. If we do not understand and call out the underlying imaginaries that underpin design motivations, any tool aiming to guide designers to the right problem to solve, or reflect on their project impacts, will fail to lead to projects that create a different kind of impact, and create change rather than continue to support a status quo. Social imaginaries describe the interpretation of potential futures shared by many people (Taylor, 2002). In paper 1 (Chapter 14) I outline my use of social imaginaries as "mental frameworks that communities use to make sense of the world by providing narratives within which people contextualize their experiences to themselves and to each other." Social imaginaries therefore have a structuring and coordinating effect on efforts to design desirable futures, but also what is considered achievable and feasible given the interpretation of reality through such imaginaries. Under this umbrella, sociotechnical imaginaries cover those imaginaries that encapsulate how people imagine specifically the connections between technology and social aspects (Jasanoff, 2015), or as Ballo (2015) puts it: “collective visions of desirable and feasible (techno-scientific) futures” - such as how technological progress is relevant for social wellbeing, how technology shapes and facilitates social relations and social possibilities. In other words, sociotechnical imaginaries cast explicit expectations and roles for technologies in regards to future social structures (Lockton *et al.*, 2019; Skjølsvold and Lindkvist, 2015).

5.1.1 Method of investigation

To investigate this question I reviewed prior research by STS and CSCW scholars, who have investigated and illustrated the formative impact of socio-technical imaginaries on the design processes underlying different technologies across various domains and fields of context from governance to healthcare, mobility, block chain and living spaces (Ballo, 2015; E.S. Ruppert, 2018; Husain *et al.*, 2020; Jasanoff and Kim, 2009; Lustig, 2019; Miller, 2020; Nardi and Kow, 2010; Schwennesen, 2019; Sadowski and Bendor, 2019).

In paper 1, I choose a selection of examples from this prior research to illustrate how specific socio-technical imaginaries are dominant factors in questions of designing digital technologies. Ames (2019) for example illustrates how the imaginary of the precocious boy leads to the failure of a project aiming to help children escape poverty through the use of computers (the project in question is called One Laptop Per Child (OLPC)). Dahlgren *et all* (2021) use the imaginary of the smart home as the pinnacle of future fully automated home comfort to highlight the definition of a household as an individualised decision making entity, which prioritises the design of hyper-personalised gadgets and hyper-targeted services over communal technology. Visions peddled by industry want to make us believe in utopian smart homes with fridges doubling as charming butlers, when at present, most of us are struggling with 30 unnecessary buttons on an unfamiliar microwave. This charisma also comes into play during the process of designing and shaping these digital technologies, curbing a balanced critical reflection upon what things are versus what they could be often already during the process of ideating. Both cases illustrate how superficially well intended tech projects make unconsidered value trade offs, depending on the imaginary that drives them - in these cases, the priorisation of individualism and autonomy over communal and social structures, for example shown in the priorisation of coding competitions over repair facilities in OLPC (Ames, 2019).

But not only products and services are shaped by the unconsidered value trade offs driven by imaginaries. Graf and Sonnberer (2020) explain that also stakeholders and potential users are configured not only through the technology being build, but also by the imaginaries that are being maintained of these stakeholders during the design process. Lockton *et all* (2019) argue that Research through Design (RtD) should contribute to creating change not only by identifying openings in behavior but first and foremost by understanding how people think. For example, rather than designing systems that encourage people to engage in more energy conscious practices, they argue that we need to address how people think about climate-change first. Building on this argument I propose that before engaging in productive and impactful value work, designers first must consider which imaginaries configure their thinking about values. Being critically aware of existing imaginaries allows designers to guide their engagement with values more intentionally.

This review of prior work in socio-technical imaginaries in the design of digital technologies resulted in a tri-partite framework for analysis of dominant imaginaries in a given design project. This framework is explained in more detail in paper 1, but I will summarize the its main components below.

5.2 Three dominant imaginaries to account for

In 14, I identify three dominant imaginaries, that I argue shape value considerations in design practice in impactful ways: imaginaries of users, imaginaries of technologies and imaginaries of design methods themselves.

5.2.1 Who? Imaginaries of Users

Designers and engineers shape user identities and expectations during the technology development process by defining and reinforcing relevant user characteristics (Woolgar, 1990; Akrich, 1992) (for example in the form of personas). While the concept of the user in general has been debated in HCI (Baumer and Brubaker, 2017), it remains strong and plenty used. Imaginaries of future users influence design processes and product outcomes by projecting idealized user behaviors and desires. For instance, in designing a smart electricity grid dashboard, future users were imagined as knowledgeable and invested, overshadowing the existing, less skilled users (Skjølsvold and Lindkvist, 2015). Similarly, autonomous driving designers assume future users will rationally embrace the technology despite current public skepticism (Graf and Sonnberger, 2020). While efforts such as human centered design or user research are meant to counter preconceived notions of users to design for, these example cases highlight how persuasive user imaginaries are. Despite better intentions, selective interpretations of user research remains in line with the expected and desired imaginary of user to be designed for.

5.2.2 What? Imaginaries of Solutions

Technological solutionism has long been critiqued within HCI, yet also remains strong in innovation efforts across both academia and industry (Gillespie, 2020; Markham, 2021; Roberge *et al.*, 2020). This solutionist approach views technology as infallible and necessary, often disregarding the complex social and emotional aspects of life (Campolo and Crawford, 2020; Alkhatib, 2021). Anticipatory imaginaries defined by such solutionist views streamline values like trustworthiness and safety into simplistic narratives suitable for technical implementation, overshadowing the nuanced trade-offs between different values (Pink, 2022; Powell, 2018). Consequently, designers are often left with using technology as the sole problem solver, limiting the design process's openness to social, political, and economic factors and excluding non-technological solutions, which could be vital components in addressing the identified problems. Designers and social scientists have extensively critiqued the deterministic tendencies in sociotechnical imaginaries, which shape technology design and use based on desirable outcomes for specific social contexts. However, they persists and are reinvigorated, particularly in AI and data-driven projects, as we will further discuss in chapter 6.

5.2.3 How? Imaginaries of the designerly process

Projecting future end-users and technical solutions through shared imaginaries creates exclusive social bubbles among designers, legitimizing certain design approaches while marginalizing others. These shared imaginaries shape work practices, favor specific methods, and reinforce existing disparities and power imbalances (Ackermann, 2023; Costanza-Chock, 2018; Irani, 2019). Business executives can become designers for a day after a design thinking seminar (Ackermann, 2023; 2021), but local activists creating communal initiatives to improve neighborhoods often cannot (Costanza-Chock, 2018; DiSalvo, 2022). The design process thus becomes a negotiation of technopolitical aspects, influenced by present sociotechnical imaginaries that determine the focus on certain issues, the perceived feasibility of solutions, and the accepted limitations.

5.3 Implications

With the presented framework I argue that changing design processes to center values is problematic without considering the sociotechnical imaginaries they are connected to. In order to understand how ways of thinking about ethics enables or precludes ethical action by designers, the three perspectives provide productive angles from which to question the significant influence of imaginaries on value work in design. The discussed projects' values for example — autonomy, efficiency, and independent learning — neglect alternative futures that emphasize communal learning, social concerns, and shared mobility. These values, shaped by sociotechnical imaginaries, become normative and dictate acceptable behaviors and roles, reinforcing socially established identities, often ignoring political dimensions like gender, race, and socio-economic status. Addressing such socio-political imaginations in technopolitical systems is crucial for analyzing normative trends. We must question who benefits from these imaginaries, who participates in and shapes them, and how they would change with broader contributions to understand how such imaginaries pre-configure the lenses that designers and developers could bring to potential value work.

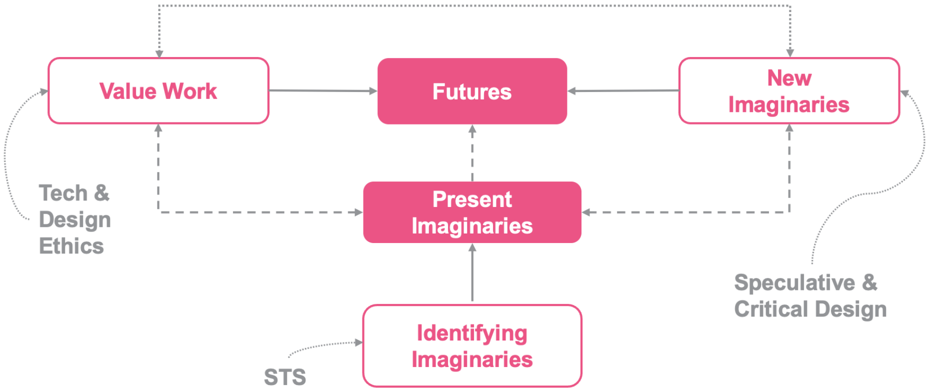


Figure 5.1: Diagram showing which areas of research work through which mode of investigation to engage with desirable futures. Tech and Design ethics as described in the introduction, mainly focus on work on value and ethics directly to address future concerns, while much speculative and critical design works through the envisioning of new imaginaries. Present imaginaries however, which also contribute strongly to the shaping of new imaginaries, are mostly investigated within the field of STS.

However, current design practices often disconnect value work from the imaginaries driving them (fig 5.1), and thereby risk the loss of alternative futures to those dictated by prevailing imaginaries without moral deliberation. In order to move forward, we need to investigate more deeply the productive entanglements between imaginaries and ethics, and develop ways of engaging such entanglements to challenge imaginaries contributing to limited ethical exploration.

6 Automating Ethics - Imaginaries of AI

As reviewed in the prior chapter, sociotechnical imaginaries act as mundane frames of future visions which shape motivations and decisions during the design process of technologies, and much prior research work particular in the field of STS has engaged with this fact. However, techno-solutionist imaginaries remain strong (Markham, 2021), and experience a particular prominence in the last years as part of the AI hype launched by the releases of MidJourney, ChatGPT and similar publicly available generative AI services. Recent releases add new fuel to the imaginaries of technologies as benevolent agents, shaping humanity for the better, and entangling in even more intrinsic ways with ideas of morality (Formosa and Ryan, 2021; Cervantes *et al.*, 2020; Fossa, 2018). Based on the implications of paper 1, and the presented outline of AI ethics in the introduction, paper 2, "AI as the final moral device" presented in chapter 15, is motivated by the question:

How do designers imagine AI to factor into ethics?

The details of the methodology, as well as the process of analysis for the resulting imaginary identified are thoroughly documented in the paper, which is currently in revision to be resubmitted to the journal Big Data and Society. Below, I will provide a short summary to explain its implications for this thesis.

6.1 Method

Paper 2 is empirically based on a participatory ethnographic study of foresight workshops focusing on AI. The workshop series resulted in a collection of scenarios, and I apply a philosophy of technology lens to tease out the expectations, assumptions and hopes that participants of the workshops express through the construction of future scenarios and their conversations around them, through qualitative analysis. As foresight is a particular strategy to investigate and explore visions of the future and understand what people desire and expect to come about, it is a useful format to identify active imaginaries that people express as they formulate such scenarios. The analysed scenarios become short hand expressions of imaginaries describing what people understand, expect and hope for the relationship between AI as an emerging technology and societal development. Further details on the method of data collection and analysis can be found in chapter 15, section 15.3.

6.1.1 Philosophy of Technology

In order to detangle these notions, philosophy of technology provides a deeper dive into the logics that have unfolded around the relationships between technology and society. Philosophers of technology have delved into the question of technology and its contributions to our lives for better and worse. Philosophy of Technology draws from fields such as philosophy, science and technology studies, sociology, anthropology, and ethics, seeking to provide a deeper understanding of how technology shapes human life and to offer insights into how we can navigate technological changes responsibly and thoughtfully. Philosophy of Technology explores the nature, impact, and significance of technology in our lives and society by analyzing how technology influences human existence, ethical considerations, and the structure of societies. I consider the contributions of Borgmann (1984), as his concept of the device paradigm delivers a useful vocabulary to inquire about both the expectations of the role of technology in our live and culture, and the dynamics that shape up around it. Most notably I have found the notion of commodification of experiences and hidden labours effective to investigate assumptions about data and neutrality in automated decision making. Borgmann provides us with a lens to analyze and discuss how the design of technology can be investigated beyond its immediate use case, and builds a contrast to the usability paradigm in discussing whether and how a technology is designed to contribute to what might be considered a good life (Fallman, 2010). In particular, Borgmanns' device paradigm provides a useful logic to recognize what kind of relationship with AI and morality we might have, and where our focus should move to instead – towards the means which are being deployed to construct visions of automated morality as an end, and towards alternative means which prioritize critical and political work on the sociopolitical aspects of AI.

6.2 AI as the final moral device

In paper 2, I explain how the device paradigm's promises appear in the hopeful future of AI solving humanity's big problems. In the qualitative data collected from the observed workshops, technology is seen as a tool to make life easier, moving from work tasks to tough ethical and political questions. Participants assume AI can gain knowledge beyond human reach by accessing "all" available data and accessing the world as it is in full and truth (Goldenfein, 2019). In this assumption, there is an understanding of data as finite, or at least, that there is a specific amount of data that provides a complete overview of any given situation, thereby providing certainty in every scenario and making moral decision making more accessible (based on the assumption that moral decision making is hindered by not knowing enough). Verbeek (2002) argues that by making means more easily available through the creation of technology, such technology could lead to a further spread and increase of application in people's lives, rather than a general reduction of consideration. Applied to the case of automated ethical decision making that might mean that moral agents which provide easier access and closer integration of ethical considerations in people's lives, thereby creating a moral device, might lead to people considering a broader array of situations and decisions as ethically relevant and deserving of moral inspection.

In this vein, the analysed scenarios acknowledge and consider the importance of human perspective and insight, and recognise that in those futures of AI guided decisions, humanity might be affected negatively, losing out on comfort and autonomy. However, even where this is recognised, the outcome of such scenarios is still a future where harmony is maintained without ongoing human effort, making politics unnecessary. Such an outcome – with AI as the final moral device - is considered desirable by participants, as they express disillusionment with current political directions, and seek better solutions than what humanity has currently achieved. There is also a fear of running out of time and lacking the skills to solve our challenges. Participants are willing to sacrifice comfort and lifestyle for what is 'right,' especially if a trustworthy source, also non-human, clearly outlines the path.

However technology can't make promises — people do, and these promises involve political and social justice issues (Borgmann, 1984). Believing AI will fix existing problems ignores the fact that a device-focused culture caused these issues (Strong and Higgs, 2010). The workshop aimed to discuss sustainability and ethics, but the outcomes imagined AI solving these problems for us. In the identified imaginary of AI as moral device, AI is expected to handle large data sets beyond human capacity, aiming to uncover a pure, rational view of reality and hence being able to act as provider of universal truth. This makes AI an unquestionable authority, creating knowledge separated from political power, shifting the focus from building ethical AI to creating machines that handle ethics and morality for us. This approach leads to a de-politisation of ethics by making AI systems seem uniquely powerful and shifting accountability from creators to the system itself. If the tool is ethical, there's no need to question its creation. This view flattens power imbalances and political tensions, assuming AI can resolve all conflicts, including ecological issues, equally. AI as a moral device suggests it can answer ethical questions better than biased and limited humans. This view separates human understanding from AI's actions, making AI seem more trustworthy and uncontestable. This creates an idealized utopia where marginalized groups must fit into the norms set by AI, reinforcing existing inequalities (Alkhatib, 2021). Without addressing the political dimensions of social structures, unified ethical approaches will maintain the status quo, making it harder to challenge (Zajko, 2021; Birhane *et al.*, 2022).

6.3 Implications

Efforts such as the observed foresight workshops create spaces for practitioners to engage in critical conversations about AI and its future role, and therefore provide an effective opening to understanding how such practitioners imagine AI and ethics to factor into one another. These initiatives intentionally bring together influential practitioners at the edges of AI expertise, broadening the understanding of the complexities involved in integrating AI systems into our social fabric. Generally, such interventions facilitate creative envisioning of the future through scenarios, demos, and prototypes, fostering the development of ethical sensitivity. However, while these efforts offer imaginative ways to consider alternatives, they can also perpetuate the existing social status quo by reinforcing established interpretations of possible futures and recycling old paradigms in new forms. As layed out in the analysis of the workshop results, the envisioned futures may seem innovative and fantastical, but they remain constrained by structures of political disengagement, prioritizing technical imagination over political imagination. It appears easier to envision powerful technologies solving human problems than to imagine people engaging in the hard work of transforming socio-political systems through political labor. The roles and functions of AI in these imagined futures highlight the participants' difficulty in breaking away from dominant deterministic views of technology shaping our social infrastructures (Markham, 2021). While these visions raise expectations of technological capabilities, they diminish the perceived need for political engagement, presenting technological progress as a comforting solution to structural societal issues (Bareis and Katzenbach, 2022).

These insights provide two main implications for the research of this thesis:

1. Imaginaries of AI champion device-driven relationship with ethics, foregrounding ease, neutrality and objectivity as desirable qualities of ethical engagement, and

2. Challenging this kind of expectation of automated morality is difficult, particularly when the forms of engagement remains future-oriented and abstract, such as in the case of the observed workshops.

7

Conclusion Part 1

As papers 1 and 2 demonstrate, ethics are embedded in imaginaries, and when ethical technology is being discussed, people draw from ideas of how the world is and the visions of futures they perceive as possible and desirable. As design and futures are inherently intertwined, we can not discuss ethics in design without considering these imaginaries and the visions of the future that are represented in design efforts. Challenging such imaginaries is difficult. They are persistent, also in projects that follow visions of social benefit and well-being, and dominant streaks such as technical solutionism resurface again and again. Purely discussing or ideating around them in futuristic scenarios is not enough to effectively challenge these deep underlying world views around what is desirable and feasible.

This brings up an important challenge for the design of technology that is more ethically aligned with how we want to live – if we can't imagine technologies differently, we can not design them differently, and if we are stuck in certain imaginaries of socio-technical relationships, we wont be able to design for other types of relationships. If we remain in the expectation of technology as a device and solution to our moral quandaries, all future-oriented interventions will reproduce the same visions, irrelevant of the values we hold dear. In order to productively challenge dominant imaginaries, and experience ethics as an ongoing human process of negotiation, we require different modes of engagement, which favour other ways of questioning going beyond discussion of hypotheticals. In addition, the intensifying discourse around AI is bringing a new intensity to this discussion, as it spawns many fantastic visions of new entanglements between morals and technology, and implies a new urgency to the existing conversations around socio-technical imaginaries in design. Current imaginaries of AI correlate morality much more deeply with technically minted features - such as data driven objectivity and universal oversight - rather than human features, such as subjective experience, social relations and emotional perceptions.

The questions proposed by paper 1 provide a first lever to highlight the potential power relations embedded in product imaginaries, which preframe how values and desirable futures become translated into technical systems and product features.

The framework and vocabulary presented in paper 2 enables designers to recognise and identify the seductive narratives and tropes active in AI imaginaries, to highlight where these imaginaries contribute to a disengagement rather than a deeper refocusing on ethical practice.

However, the question remains of how to shift these imaginations in ways that center new ways of interacting with ethics. How could modes of relating and experiencing ethical tensions in addition to imagining look like? And how would such modes of re-thinking affect actual practice?

Part II

Ethics IRL

8 From Theory to Practice

Investigating the underlying imaginaries helps in understanding how various ethical decision making efforts are being motivated and legitimised. It does not however by itself help us address the issue of ethics in practice when developing technologies. When investigating how to create spaces to challenge imaginaries in order to get to "ethics" in practice, we can not ignore the large amount of work that has been done to create different ethical practices and ensure value alignment and ethical conduct in applied contexts. As mentioned in the introduction, there has been an explosion of efforts particularly in the field of AI (Khan *et al.*, 2022; Jobin *et al.*, 2019; Corrêa *et al.*, 2023), but also in the area of design practice throughout the last 5 years, with a particular focus on design processes and practices (Vesti *et al.*, 2024; Chivukula *et al.*, 2022). Such efforts can be roughly grouped in two buckets - efforts that provide prescriptive materials, such as codes of conduct, value manifestos and checklists which are aiming for compliance, and materials meant to expand the technical mindset through reflection.

8.1 Ethics as compliance

As elaborated in chapter 4, a vast collection of principle and value sets has been developed across academia and industry to inform the development of technologies (Khan *et al.*, 2022; Fjeld *et al.*, 2020; Floridi *et al.*, 2018; Poel, 2016). In a review of global AI ethics guidelines, Jobin et all (2019) report an emerging convergence around five main principles: transparency, justice and fairness, non-maleficence, responsibility and privacy. However, they also report on widely varying interpretations and justifications for these principles. Researchers reviewing the impact of prescriptive modes however share the consensus that they are not producing the desired impact in addressing the harms they aim to remedy (Hagendorff, 2020; Mittelstadt, 2019; McNamara *et al.*, 2018). One reason for this lack of impact is the gap between the prescribed normative principles and any link to practical integration of them (Fjeld *et al.*, 2020; Sadek *et al.*, 2024; Corrêa *et al.*, 2023). These prescriptive modes are being introduced into a field of industry that is largely used to considering ethics outside of their field of responsibility, and making prescriptive modes impactful in practice will require a larger effort than many of the more high level propositions seem to acknowledge (Munn, 2022). In fact, as Mittelstadt (2019; 2023) argues, not considering this gap carefully can lead to ethics washing by influential coorparations: presenting a list of promising principles for conduct, which give the appearance of ethical responsibility without making substantive changes to their practices. This strategy covers normative disagreements and political in-action, making these prescriptive methods more of a distraction than ways forward.

In addition to principles and values, a large collection of technical tools have been developed to evaluate biases (Hort *et al.*, 2024), fairness metrics (Mehrabi *et al.*, 2021), and other quantified measures as means to assess the ethical performance of systems, both in academia and in industry, and in collaborations. Examples are the "Everyday Ethics for AI" resources by IBM (IBM, 2022), which provides a mix of guidelines and technical tools, Microsoft (Young *et al.*, 2020; Microsoft, 2024) and Google (Google, 2024; Pair, 2024) host a collection of tools developed by academics and engineers, and even PwC provides a Responsible AI toolkit (2024). The tendency of such tools is to provide evaluative checkpoints for technologists and designers (Wong *et al.*, 2020), to ensure that certain thresholds of what is considered ethical are upheld (Calvi and Kotzinos, 2023). Such tools reward compliance to often quantitative measures or procedural mile stones, such as running risk assessments, quantitative metrics of fairness or bias performance, or assessment procedures at certain points of the process.

However, such technical tools fail to produce the ethical results they promise on two accounts. First, existing tools and methods aiming to close the gap between principles and technical practice are ineffective as they are either too flexible, and thus vulnerable to ethics washing, or too strict and thereby unresponsive to context (Morley *et al.*, 2020; Whittlestone *et al.*, 2019). Second, such tools only tend to technical parameters (Binns *et al.*, 2018), and do not take into account the socio-political undercurrents of moral aspects such as fairness (Friedler *et al.*, 2016; Corbett-Davies *et al.*, 2017; Hoffmann, 2019; Bennett and Keyes, 2020), bias (Richardson, 2021), justice (Davis *et al.*, 2021; Green, 2021a) and others, and reduce them to mathematical and hence measurable qualities (Zajko, 2021; Andrus *et al.*, 2021). In the worst case, such formats of ethics are adopted as performative acts, which limit the actual discussion space around potential issues rather than opening it up (Roberge *et al.*, 2020). They hence create distractions from existing issues, rather than creating authentic attempts for change (Phan *et al.*, 2021).

8.2 Ethics as Reflection

Another group of ethics tools aims to enable a broadening of world view, and encourage technologists to widen the field of what they consider relevant for ethical consideration in their work practice. The primary activity of such tools and interventions is reflection on new or different perspectives, and the primary outcome desired is a heightened ethical awareness or sensitivity. Van de Poel and Gorp (2006) or Steen (2015) for example recommend increased modes of reflection and consideration between designers and engineers, aiming to encourage broader thinking through the contexts and values present. Frauenberger *et al.* (2017) argue that work in HCI has become more dynamic, unpredictable and participatory and requires a change in approach to ethics to reach into these changed processes and practices beyond the anticipatory approaches to ethics traditionally deployed. They suggest to address these changed needs by establishing a shared *ethos* for a project, a shared ethical sensibility which guides actions in context and changes the question from whether principles have been followed to whether the right thing has been done given the circumstances. Suggested operations are a kick off of a project that focuses on establishing shared ethos, a facilitator that helps the coherence of the ethos along the way and ongoing documentation of ethos related reflections and decisions made according to the ethos. As mentioned in chapter 8, Value Sensitive Design aims to connect inquiry into values with practical integration of such values into the design process, by providing a large collection of methods and tools. However, VSD has faced some criticism in regards to its limit of value discovery during the engagement with affected communities (Le Dantec *et al.*, 2009), while leaving the process open for any kind of values to be centred (Borning and Muller, 2012) - also non-ethical values, such as usability, or efficiency (Albrechtslund, 2007). And despite the extensive development, there continues to be a lack of guidance for implementation of VSD methods and tools in line with the intention of the original authors, and adoption within broader industry circles is low (Winkler and Spiekermann, 2021; Rotondo and Freier, 2010; Manders-Huits, 2011).

Many other smaller initiatives aim to fashion ethical considerations into the design process in modular fashions, in the form of workshops, toolkits, and other process modifications, created both within academia and in by industry. These materials recommend reflection and provide methods and exercises for different steps of the proposed processes (Zou, 2018; OPEN, 2021; DDC, 2021; Thoughtworks, 2021; Dot.Everyone, 2024; Tangible, 2024; Gispen, 2017).

However, the multitude of publications mainly highlight that so far, they have failed to create a consensus within design practice on how to reliably address the ethical challenges posed by rapid technological progress. Where some of these efforts are put into practice, it happens on a niche scale by individual labs and studios applying their own processes, and aside from fueling debate within the design community. Not many of these have developed into more than conversation pieces for design conferences and events. For all these efforts within these fields however, their impact appears to be minimal.

Designing for Care

9

As elaborated in chapter 4, many design methods have been developed to reflect on values, to guide consequence scanning (aligned with consequentialism), or define principles for certain conduct. While many such efforts claim success in increasing moral awareness and reflection amongst the participants, quite a few of the designers of such efforts also highlight the shortcomings of their efforts in regards to affecting deeper change (Luján Escalante *et al.*, 2022; Harmon *et al.*, 2016; Madaio *et al.*, 2020). Few methods accommodate a more relational viewpoint on ethics—one that does not merely use scenarios to derive generalizable rules of conduct. This observation motivates paper 3 (chapter 16), titled "Sensing Care Through Design: A Speculative Role-play Approach to "Living with" Sensor-supported Care Networks", which has been published and presented at DIS 2023. It poses the question:

How can we design formats of engagement for practitioners to work through the ethics of imagined tech futures?

To move ethical deliberation beyond mere compliance and reflection, and to engage with the sociotechnical imaginaries that shape our envisioned technological futures, design interventions need to give space for the contextual and relational aspects of ethics in a more embodied manner. Paper 3 reports on a method that does not aim to deliver solutions or generalizable insights, but instead creates a space where imaginaries can be contextualised and challenged, fostering a different mode of ethical thinking. Drawing from the ethics of care, we propose an alternative approach that views ethics as a situated and relational process of iterative engagement with each other's needs. This method focuses on how we engage with ethical dilemmas, rather than creating prescriptive rules to resolve them. By putting ethical considerations into context, we create challenges and tensions that encourage different ways of thinking.

9.1 Care in HCI

Care as a concept for designing interactions between people and technologies has received more attention in HCI recently (Key *et al.*, 2021; Light and Akama, 2014), as digital systems become more entangled with social relations (Frauenberger, 2020), and imaginaries of various autonomous systems are construed around the idea of technologies taking care of certain aspects of life: a smart home taking care of certain parts of house work (Key *et al.*, 2021), a smart city taking care of infrastructure maintenance (Forlano and Mathew, 2014), or health technologies enabling different kinds of care for physical (and sometimes also mental) wellbeing (Wynsberghe, 2017).

Prior discussed principle-based frameworks aiming to protect principles like privacy, independence, and autonomy, center people as independent individuals, who are capable of always making rational and informed decisions about technology use. This perspective places an unrealistic burden on individuals to make technology-related decisions (Dourish, 2004; Seberger *et al.*, 2021). The logic of care recognizes that people are interconnected, entangled in networks of varying needs, relying on each other for decision-making, and embedded in time (Mol, 2008). It acknowledges that decisions around technologies and their use are influenced by conditions that are not ideal, potentially emotionally charged and depended on the social relations they are embedded within (Groves, 2015). The concept of care has been applied through design experimentation in various contexts within HCI (Light, 2021; Light and Akama, 2014; DiSalvo, 2012; DiSalvo, 2022) to investigate pathways for designing technologies outside of individualistic and utilitarian motives.

Despite the growing focus on care in HCI, defining the term "care" remains challenging in the context of relations between humans and technologies. For one, care is inherently political (Woodly *et al.*, 2021). To care means to cast judgments about whom to care for, how, and to what extent (Mol *et al.*, 2010). Particular in sociotechnical systems, care can quickly slip into paternalistic patterns (Martin *et al.*, 2015; Groves, 2015) when having to deal with tensions and conflicts of prioritisation and appropriateness of social behaviours (Yacchirema *et al.*, 2017). In addition, expecting algorithmic systems to care requests that they can accommodate uncertainty, and pay considerate attention to tensions and conflicts in their designs and interaction patterns (Light and Akama, 2014). However, these systems can only simulate consideration, despite their potential anthropomorphized appearance (Jones *et al.*, 2023). In other words, while these systems can be designed more care-fully (Light, 2021), building systems that care remains impossible.

9.2 Developing the Method

The study for paper 3 was initiated as part of the DCODE collaboration between PhD students and non-academic research partners, as part of the first prototeam cycle (as mentioned in chapter 3).

Our prototeam collaborated with two organizations specializing in sensor-based technical systems. The Institute for Amsterdam Metropolitan Solutions (AMS) in Amsterdam develops sensor networks to optimize city operations, such as mobility flows, energy infrastructures, and asset management. In Edinburgh, the Advanced Care Research Centre (ACRC) aims to integrate machine vision sensors into the homes of elderly residents requiring care, promoting independent living. Both organizations sought innovative methods to explore trust and privacy through forward-looking scenarios and user-centered explorations. To understand their needs, we conducted fifteen in-depth interviews over two months with experts from AMS and ACRC. Our interviewees, chosen for their diverse perspectives on sensing networks, included citizen scientists, interaction designers, technologists, and city legislators at AMS, and healthcare practitioners, nursing scholars, social scientists, and technologists at ACRC.

The first five co-authors reviewed and discussed all interviews, using an iterative, thematic analysis approach. This method ensured that each new interview built on the insights from previous ones. Our timeline was grounded in issues and concerns raised during industry partner interviews, highlighting recurring themes across both organizations: technological development, politics, business, and healthcare. We gathered significant concerns, predictions, and assumptions on these topics, framed them as individual news items, and linked them into a coherent, progressive narrative.

In response, we developed a workshop method that utilises speculative role play, drawing from speculative enactments (Elsden *et al.*, 2017), critical play (H. Tan *et al.*, 2022) and Participatory Design Fiction (Prado de O. Martins and Oliveira, 2017), which is reported on in detail in paper 3. The method uses a fictional narrative represented as a timeline of speculative news items, inspired by *2038 The New Serenity* 2022 and Wong and Nguyen, 2021, and a card deck that guides the creation of role play scenarios. Our motivation was to create a framing that attends to four primary concerns deemed important for a method of ethical engagement, that stretches beyond static value alignment and accounts for situated and contextual value tensions, in particular in situations where people encounter vulnerability and dependency, making them incapable to contest. The four concerns are:

- Multiple voices and actors
- Interdependence

- Situatedness
- Performativity

The details of the application can be read in paper 3, together with the variations we deployed and the reactions of our collaborators.

9.3 Reflecting on Speculative Roleplay

Paper 3 addresses industry partner concerns regarding the challenges posed by developing sensor network technologies by re-imagining living with these technologies from a place of care. The presented method illustrates that providing care involves more than just measuring vitals or recording situations; it requires nuanced reactions and the negotiation of value tensions. The workshops supported participants in experiencing and reflecting on these tensions, illustrating that ethics and human values are complex, re-workable, and time-dependent. This approach helps designers move away from user-centered optimization towards facilitation and making space for care. Our outcomes indicated that designing sensor network-based systems must go beyond compliance with privacy policies and consider trust and other values as shifting components shaped by living. The method demonstrated the importance of structured speculation with both design teams and stakeholders, emphasizing that designers need to experience these value tensions firsthand. The improvisational role-playing nature of our method brought value tensions to the forefront, showing how different positions, priorities, and values interact and come into tension.

The combination of two familiar approaches - speculative futures and role play - in the setting of concrete care scenarios presents a novel orientation for such a research through design approach. In addition, our implementation of the method involved both potential users and technologists in the same context, to connect potential needs and perspectives on lived experiences with technical expectations and imaginaries of potentially possible technology (Stilgoe *et al.*, 2013; Yacchirema *et al.*, 2017). Most importantly, we aimed to create a space where different perspectives can come together, and value tensions can be experienced in a more consequential yet safe space, rather than discussed on hypotheticals (Seberger *et al.*, 2021; Medler and Magerko, 2010).

Finally, we designed a flexible workshop setup that allowed participants to role-play future scenarios without extensive preparation, catering to different content, participant types, and group sizes. This flexibility enabled diverse audiences, including designers and non-designers, to engage with the complexities of socio-technical sensor networks through storytelling and play (Elsden *et al.*, 2022; Yacchirema *et al.*, 2017). Tapping into the embodiment aspect of the role play shifted the experience of the emerging ethical concerns from something abstract to contextual encounters that became more tangible in the affordances they provided for the participants.

However, while the method helps researchers, interdisciplinary design and development teams, and potentially affected users, to explore tensions around care relations and technologies in a speculative manner, limitations remain. The set up of the workshop requires well versed facilitators, that are focus on the support of the role play rather than the arising conversations, and that provide enough scaffolding for participants to stick within their improvised scenarios when the unfolding narrative becomes potentially convoluted. While we ran the workshop four times with different groups, and closely collaborated with our non-academic partners, we did not perform the final workshop format with industry professionals, but only with professionals within research contexts. This might have contributed to the openness of participants to debate the emerging conflicts proactively instead of defensively, and acknowledge the complexity of the created situations rather than discount them.

9.4 Implications

While the accommodation of multiple voices, interdependence, situatedness and performativity through the role-play provides a more experiential and tension oriented approach to negotiating value trade offs in the design of sensor-driven actors, our proposed method remains speculative - the encountered dilemmas are created from a relevant stack of topics, but the participants might still only encounter a fleeting investment. In addition, while the workshop brought out the affective dimensions of encountering concrete situations of value tensions and putting participants on the spot for resolution, the resolution remained hypothetical and without impact beyond the framing of the workshop. Such a method is then potentially effective in raising ethical awareness and making the relational and situational nature of ethics more tangible for participants. But it also leaves them with open questions as to how to pro-actively import such a change in world view into their ongoing practice.

Many interventions and approaches to more ethically reflective work practices stop at this point - recognising the limitations and challenges in bridging impact of such methods beyond bracketed sessions and speaking to the importance of organisational structures and pressures to enable participants to carry their insights and learning into their everyday. But these challenges are not trivial, and are currently falling somewhere between the cracks of designing practices and methods and adopting them in fields of business management. It is in this connective step where many interventions lose traction. The next step in my research is therefore investing this connective part.

10 Doing Ethics

Given the ongoing uncertainty and fickleness in the ethical dimensions of tech design, researchers are turning towards studying existing practice to investigate what influences professionals' engagement with ethical tensions in their everyday work life, beyond provided tools and guidelines. Normative ethical inquiry is very different than addressing ethical dilemmas in practice (Munn, 2022) and significant research has been done to investigate where the points of tension lie between expected adoption and actual implementation of ethical toolkits, such as described in chapter 8. This research has been mostly done through interviews and workshops with companies and practitioners, both design and development-oriented, to understand how they conceptualize ethics (Dindler *et al.*, 2022), how interventions should be designed to align with practice (Chivukula *et al.*, 2020), and how they address ethical concerns when working with data (Taylor and Dencik, 2020) and developing AI (Ibáñez and Olmeda, 2022; Yildirim *et al.*, 2023; Vakkuri *et al.*, 2020). Researchers have also identified the gaps and challenges encountered by (Liao *et al.*, 2020) and examined how practitioners engage with existing tools (Yildirim *et al.*, 2023). Chivukula (2021) reviews 63 ethics-oriented design *methods*, while Wong (2023) specifically reviews the impact and expectations of *toolkits*, with both highlighting the mismatch between theoretical and practical ethics work. Much of this research highlights the failing of analyzed tools and processes to account for the environments where such practices are situated. Metcalf (2019) reported that industry leaders see ethical efforts as being in tension with industry structures like technological solutionism and market fundamentalism. Lindberg (2024) found that leaders consider the lack of awareness and practices that align with market demands as significant challenges.

However, few researchers have dug into the challenges reported by professionals for longer timeframes and deeper integration. While studies using interviews and workshop sessions provide valuable insight into frustrations that professionals are aware of, they are unable to observe the embedded and unarticulated coping dynamics, professionals engage in, but might not necessarily perceive as relevant to the studies focus, or take certain limitations and unspoken rules for granted and not relevant to mention. For example, Shilton's (2013) conception of Value Levers describes how organizational and operational structures in design teams and labs influence the way values and ethical deliberations are incorporated into daily routines. She identifies various infrastructural aspects that support and embed discussions about values in technology design, noting that in a market-driven design field, the constant pressure for technical innovation often hinders a slow and deliberate value-driven design process. More recently, Lindberg (2023) conducted a 10-month field study to co-design materials that sensitize design practitioners and reported on organizational barriers, such as the difficulty in gaining both buy-in and motivation, and logistical barriers like time and resources needed to develop ethics interventions that go beyond "tick-box exercises."

Such research mirrors my own experience, that it takes much more than well developed tools and methods to create change in practice, and make the impact of such interventions long lasting. Carrying insights, reflections and potential mind shifts beyond the frames of interventions such as workshops, seminars, or design sessions, requires participants to affect more than themselves and recognise how they practice ethics across their entire work day, not only in the confined boxes of ethics tools and sessions. Paper 4 (chapter 17), "The Affective Experience of Ethics", turns towards my last research question:

How do practitioners experience the doing of ethics in an organizational everyday context, and how does that affect ethical practice?

Paper 3 and the "Sensing Care" workshop already provides a suggested alternative on how to engage with ethics beyond checkboxes and reflections, and has provided valuable insights into the tensions that come up when imaginaries are being questioned and trade offs need to be negotiated. What is missing is the insight about the connecting tissue, the work that happens outside of and as a result of such clearly defined and timely bounded ethics sessions, across relations and interpersonal politics. Research has already highlighted what needs to be in place in order for toolkits and interventions to actually be accepted and adopted into practice. This research matches my experience as well, with specific team dynamics and relations necessary to bin in place in order for productive consideration of ethics with the intent of change. But what happens when these things are given?

10.1 Ethical Sensitivity and Moral Stress

As pointed out in chapter 8, one common aim for practical ethics interventions is the development of stronger ethical awareness, and deeper reflection. However, few tools actually define what they mean with these terms. Gray and Chivukula (2019) have explored the cultivation of ethical sensibilities through the concept of mediation, and Pillai *et al.* (2022) have investigated the importance of social connection for navigating ethical tensions, suggesting that the emotional context of ethics may be an important consideration for bridging the gap between principles and practice (Dunbar, 2005).

While emotions have received more attention within HCI with the turn towards the third wave (Bødker, 2015; Wadley *et al.*, 2022), the majority of the work has put its focus on the emotional experiences of users of technology (McCarthy and Wright, 2005; Mark *et al.*, 2015), and taking the importance of emotional well-being of consumers into moral account (Calvo and Peters, 2014; Mark *et al.*, 2015). HCI has turned only recently towards the importance of emotional experiences of the practitioners designing technologies, such as through Soma Design (Höök, 2018). Soma design has provided aspects of focusing on affective bodily experiences for designers to be inspired and think differently about the technologies they design. This provides an interesting and generative opening into integrating other aspects of understanding positive experiences beyond utility or innovation driven narratives into the design of technologies. Garret *et al.* (2023) and Popova *et al.* (2022) have applied the lens of soma design to demonstrate that connecting ethical sensitivity to emotional experience provides a generative approach to understanding the embodied perspective of ethical decision making.

Outside of HCI, disciplines that have a longer history with investigating ethical dimensions in practice have mapped relevant concepts around the tensions professionals experience in more detail. Research on ethical sensitivity (Weaver, 2007; Widder *et al.*, 2023; Wong, 2021) for example recognizes that practitioners already are ethically engaged, and aims to understand when and how ethical decisions are being made. Boyd and Shilton (2021) have reviewed similar research and apply the importance of Ethical Sensitivity for HCI. Research such as by Boyd and Shilton (2021) and Garret *et al.* (2020) design interventions facilitating ethical reflection indicate that an increase in ethical sensitivity and moral awareness is considered a *positive thing* for everyone involved.

However, as research from other fields shows, ethical sensitivity and moral awareness can also have negative consequences for the ones acquiring it (Weaver, 2007). Ethical considerations are emotional and embodied (Su *et al.*, 2021; Popova *et al.*, 2022; Garrett *et al.*, 2020) especially where these challenge existing social and political norms in a corporate context, coming into conflict with the techno-optimism many companies establish in their culture of work (Gould, 2009). In moral psychology, an increase of self-reflective behaviour is considered the outcome of higher moral awareness, a concept that has been well researched from biological, psychological and socio-cultural viewpoints (Jordan, 2009; Reynolds and Miller, 2015). Studies of moral awareness have shown that along with an increase of self-reflective behaviour, stronger moral awareness also makes people more self-conscious and uncertain about their own actions (Reynolds and Miller, 2015). This uncertainty and self-consciousness comes into play when a person becomes aware of the fact that a moral decision they are making is requiring to act a certain way, but are prevented from following through on this action because of circumstances beyond their control (Reynolds *et al.*, 2012). Such circumstances might be norms that conflict with the action in question, missing opportunity to do the action, or missing resources or capabilities.

This discrepancy has for example been studied in management studies (Reynolds *et al.*, 2012), in cases where an individuals and an organizations goals or opinions about the right choice of action don't align. Where such conflicts occur, the people involved experience moral stress - a psychological state associated with negative emotions of uncertainty and inadequacy. However, moral stress does not only occur as a result of a moral decision - action mismatch. It can also occur when a person feels they can only choose between two justifiable but conflicting options, where both could be considered morally right, or two options where both are perceived as morally wrong (Reynolds *et al.*, 2012). Nursing studies have debated this emotional experience under the concept of moral distress, as a result of too high workload, time pressures and role ambiguities, amongst others. Such circumstances put nurses under pressure to adapt their care to external logistical demands, rather than follow what they consider the morally right to do in order to provide good care. Such moral distress puts a high emotional strain on nurses beyond their already intense work structures, one that is importantly often not taken into account as being a factor of exhaustion. Where moral distress is not acknowledged as negative and draining, nurses react with limiting the efforts they put into patient care, avoidance of or failure to act in morally difficult situations (Reynolds *et al.*, 2012), and also dropping out of the occupation (Morley *et al.*, 2019).

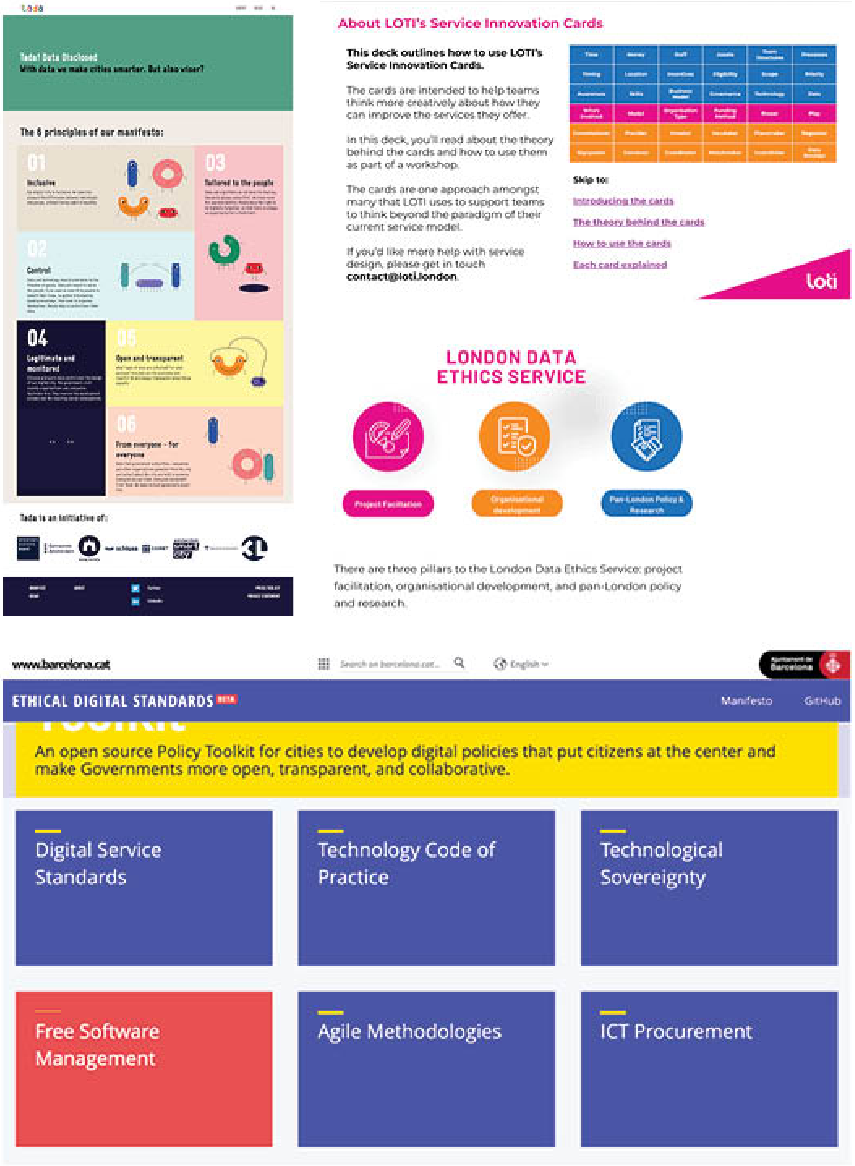
10.2 Research Context

The rapid digitalization of the public sector has sparked a debate in Europe about the implications and ethical challenges of smart data-driven technologies, particularly their impact on the public domain (Thylstrup, 2019; Floridi, 2014). In response, several major European cities have developed codes of conduct, principles, value sets, and manifestos to guide the future development of data-driven systems. Amsterdam, for instance, has created the prominent Tada Manifesto in collaboration with citizens, businesses, and governmental bodies, integrating its principles into city operations through specially designed workshops (*Tada.city* 2023). The London Office of Technology and Innovation, a coalition of the GLA, the London Councils, and the

London borough councils, has established the London Data Ethics Service, which provides project facilitation, organizational development, and pan-London policy and research through in-house experts (City of London, 2024). Barcelona has launched the Ethical Digital Standards initiative, encompassing digital service standards, a Technology Code of Conduct, methodologies for agile development, and the Ethical Digital Standard Manifesto, which articulates the city's vision of technological sovereignty and digital rights (2023). See 10.1 for reference of the mentioned materials.

In the study reported in paper 4, I collaborate with a major European City, which, for anonymity reasons will be referred to in the following as The City. The City responded to the mentioned debate by bringing together local authorities, businesses, knowledge institutions, organizations, and residents of The City to provide a direction for how data driven technologies should be designed in alignment with a shared set of values, towards an ethical, responsible, and inclusive smart city. The result of this effort was the development of a manifesto (referred to in the following The Manifesto) for responsible data use and an accompanying toolkit (referred to in the following The Toolkit).

The Manifesto addresses the "big questions of the digital revolution" and calls for the formation of a community tasked with making The City a moral thought leader and exemplar in tackling these significant issues. To integrate the Manifesto into everyday development efforts, the municipality of The City explored how civil servants could apply its principles in all data-driven projects. Recognizing that effective integration requires time and resources, the municipality developed the Toolkit to provide these necessary supports. While the Manifesto targets all stakeholders, the Toolkit is designed specifically for internal use within the municipality's organizational structures and employees. Over five years, the Toolkit was crafted as a workshop format to help teams understand and connect ethical values to their daily work. Teams participate in an introductory workshop led by an ethics expert or trainer to learn about the Manifesto's values and their relevance to their work. Following this workshop, each team appoints an Ethics Guardian and receives materials to conduct smaller workshops for individual projects.



tada

Tada! Data Disclosed
With data we make cities smarter. But also wiser?

The 6 principles of our manifesto:

01 Inclusive

02 Control

03 Tailored to the people

04 Legitimate and monitored

05 Open and transparent

06 From everyone - for everyone

Tada is an initiative of:

About LOTI's Service Innovation Cards

This deck outlines how to use LOTI's Service Innovation Cards.

The cards are intended to help teams think more creatively about how they can improve the services they offer.

In this deck, you'll read about the theory behind the cards and how to use them as part of a workshop.

The cards are one approach amongst many that LOTI uses to support teams to think beyond the paradigm of their current service model.

If you'd like more help with service design, please get in touch **contact@loti.london**.

Skip to:

Introducing the cards

The theory behind the cards

How to use the cards

Each card explained

loti

LONDON DATA ETHICS SERVICE

Project Facilitation

Organisational development

Pan-London Policy & Research

There are three pillars to the London Data Ethics Service: project facilitation, organisational development, and pan-London policy and research.

www.barcelona.cat

Search on barcelona.cat... English

Ajuntament de Barcelona

ETHICAL DIGITAL STANDARDS BETA

Manifesto GitHub

An open source Policy Toolkit for cities to develop digital policies that put citizens at the center and make Governments more open, transparent, and collaborative.

Digital Service Standards	Technology Code of Practice	Technological Sovereignty
Free Software Management	Agile Methodologies	ICT Procurement

Figure 10.1: An overview of the publicly available materials from the three mentioned example cases of Cities developing their own responsible data and development resources

10.3 Ethnographic Method

In order to better understand the affective experiences and peoples reaction to them, I want to witness their actions and behaviours as they engage with situations and everyday dilemmas of their work, within ethical toolkits and beyond. Rather than turning towards the creation of more methods, tools and interventions, I wanted to observe the efforts that are already existing between the margins of concrete ethics-oriented materials and everyday design decisions. With this project I followed what Pink (2021) termed my ethnographic hunch. From both from my personal experience prior to the PhD in various design and development teams, as well as my readings of existing academic research on challenges of adoption of ethics tools, I wanted to investigate these challenges as they presented themselves in existing tools and organisational structures, not from my intervention through close-up involvement in some form of participative role, in the everyday setting these tools are embedded in (Stewart, 1998). In contrast to the Sensing Care project discussed in the previous chapter, I did not want to present something I had developed as an outside expert, where I am the designated voice to define how things should be done. Rather I intended to observe and learn from what had been created from within organisational culture and working infrastructures to address the perceived challenges and issues. In explaining to me how things are expected to work, specific assumptions and perceptions are expressed about what the problems are and what desirable ways to address them look like. This communicates more about the relations and roles of different teams and actors in the organisation than if I had observed a situation, come up with an intervention and brought that back to the teams.

For this purpose I conducted ethnographic work with a team of AI developers at the municipality of a major European city with a long standing investment in creating responsible smart city technology. The details of my engagement can be read further in paper 4. I opted for ethnographic methods as the most appropriate qualitative research technique to study development teams and communities by immersing myself in the daily lives of the technologists confronted with their ethical responsibilities and encouraged to deploy ethics tools in their project work. This approach involved participant observation, and participation in their daily work, to observe their behaviors and interactions. Additionally, I conducted in-depth interviews to gain a comprehensive understanding of the social dynamics and relational structures at play. In this project, my background in design industry was very valuable as I was able to contribute to the teams projects with valuable skills myself, and become a short term part of the shared development efforts. My prior familiarity with project management structures, vocabulary and tech project norms allowed me to gather a detailed and nuanced portrayal of the team, capturing the complexities and subtleties of their lived experiences. This method emphasizes the importance of context, allowing me to investigate how technologists as individuals navigate their social worlds and make meaning of their experiences with ethical toolkits aiming to produce responsible technology.

During the fieldwork, in conversation with both the teams I worked with and in reflective reporting to my co-authors, I developed a sensitivity towards the emotional reactions that team members expressed in their engagement with the tools, as well as outside of such sessions, and where these emotions re-surfaced in (at first seemingly) unlikely places around disconnected work (Pink, 2021). For example, in one meeting the project manager requested all team members to have their cameras switched on. One lead developer declined, stating that they didn't feel like it and shouldn't be made to turn on the camera if they don't want to. This created a tense atmosphere even through the screen, as other team members were hesitant to switch their cameras on or off, as that would indicate an implied siding with one person over the other. When it was my turn, I kept my camera switched on and mentioned that I prefer to have my camera on, which visibly made the project manager relax - they perceived to have my support within an interpersonal conflict. However, this conflict was not actually about the camera - it was an expression of ongoing tensions around roles, responsibilities and authority, expressed in small, everyday interactions. Such everyday interpersonal politics are difficult to capture and understand when one is not part of them, and relies purely on the report and descriptions of the people one collaborates with.

I documented my observations in extensive paper journaling and twice-weekly reports to my co-authors, working through my perceptions of them as I noted them, and bringing my hunches back into conversation with the teams. In this way, a back and fourth developed between the teams, our conversations, my participation in their work, and my reflection upon that, and the semi-developed reports I send over email. My 'digestion' of my experiences was very much shaped and influenced by the email correspondence described by Cerwonka and Malkki (2008), and a similar interpretive approach to the knowledge derived from it. Cerwonka describes this approach to ethnography as sensitive to "the closeness between emotions and ethics, and illustrates moments in fieldwork that are not only tense social situations, inevitable in social research, but also ethical quandaries that are signaled in the body." (2008, p.2) in her essay termed "Nervous Conditions" (a title very relatable for my ethnographic experience). Being integrated within the team not only as observer but also as contributor opened my research up to experience of social interactions, both positive and tense, not only from the report of the teams but also as they are signaled in the body, allowing me to immerse myself deeply enough to pick up on affective dimensions of the relational work attached to the implementation of ethics toolkits.

10.4 The Affective Experience of Ethics

In paper 4, "The Affective Experience of Ethics", I present a case study resulting from the prior ethnographic work, that highlights the emotional and relational labor that accompanies the work required by technologists. These technologists are committed to take accountability and responsibility for the ethical dimensions of their work, and encounter moral stress because of it.

Using different vignettes, the paper describes how this labor is shaped along different relational configurations (see fig 10.2), which put varying demands on the practitioners to navigate social structures, and result in different facets of moral stress:

Within team describes the relational labor the team invests into each other in order to make the practice of ethical reflection possible. Here, the observed team experienced moral stress due to a reconfiguration of their individual roles and responsibilities, which require them to be open and vulnerable in order to engage in honest self-evaluation, and to embrace the change from purely technical creator to moral contributors. Both aspects create uncomfortable emotions, and especially the later one is characterised by the shouldering of heavy responsibilities.

Beyond team describes the tensions the team experience in relation to their position within the organization, the expectations that are put on them by others, and the responsibilities they take on towards other teams. As part of the innovation department, the team was used to working with exploratory projects, but the AI hype escalated the demands and expectations the team faces from other departments and collaborators. In this case the team finds itself as moral gate keepers, who navigate mismatched expectations between ethical quality assurance and technical novelty. At the same time, the team experiences tensions around power mandates - being in the position of civil servants, they are by default not empowered to take political decisions, yet as technical experts are expected to make ethical recommendations.

Beyond organization describes the tensions and concerns that stretch outside of their organization and touches on the relationship of the team as civil servants with citizens. Particular in regards to transparency and accountability, the team experiences ambiguity resulting in moral stress. The team itself expect themselves to be transparent towards externals, yet experience feelings of nakedness and exposure, which they need to cope with individually, despite acting as representative of an institution. In regards to accountability, the team runs into challenges on how to interpret a given status quo they encounter in their problem definition. Considering a broader scope of inquiry to take into account social dimensions of innovation, the team recognises that technical innovation can only partly resolve the encountered issues, which are inherently embedded in public infrastructure.

10.5 Beyond Paper 4

The analysis of moral stress presented in paper 4 only addresses parts of the insights gathered during this research. Further themes connected to moral stress and social limitations of ethics demands were identified during the analysis of the qualitative

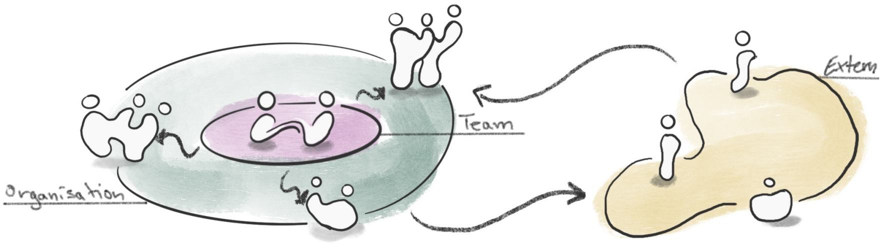


Figure 10.2: Illustrative visualisation of the relational constellations around which moral stress unfolds.

data. These themes have not yet been developed in detail, but are planned to be covered in future publications. I briefly report on them here to add to the implications of this study for the overarching intention of my thesis:

10.5.1 Downstream Ethics

One aspect of the team encountering tensions is the misfit between established project management structures and the unfinished nature of ethics. As outlined in the opening of this chapter, ethics tools and interventions generally don't seem to take into account the practical requirements of work-life in tech and design roles in industry, leaving the practitioners discouraged from actually attempting to implement them. Part of that is, as other researchers have pointed out (Lindberg *et al.*, 2023; Metcalf *et al.*, 2019), a cultural clash between established narratives of innovation and disruption. However, even where office cultures are generally embracing questions, doubts, and taking time to discuss dilemmas, there are aspects of established project management structures, such as agile [1] and scrum methodologies [2] for management of tasks allocated among the team. Agile and scrum have been adopted as project planning and execution strategies in software developement as an alternative to waterfall strategies, where project tasks and deliverables are planned in the beginning for the entire project lifespan. In comparison to waterfall planning, agile and scrum strategies emphasise iterative, small project steps, that are supposed to make the accommodation of changes and adjustments easier along the way. Both agile and scrum originated in the field of software development man-

[1] Agile is a project management and software development methodology characterized by flexible and iterative progress. It emphasizes delivering small, incremental updates rather than a complete product at the end of the project, allowing teams to respond quickly to changing requirements and feedback (Beck *et al.*, 2013). Agile practices include regular team meetings, continuous planning, and stakeholder involvement throughout the project lifecycle (Cohn, 2010)

[2] Scrum is an agile framework for managing complex projects, often used in software development. Scrum consists of defined roles, such as the Product Owner, Scrum Master, and Development Team, and employs practices like time-limited project cycles called sprints, daily stand-ups, and sprint reviews (Sutherland and Schwaber, 2024).

agement. With many designers in industry being integrated or collaborating closely with developers, the design process has been included in tasks to be managed in this manner, particular using elements of design thinking. This has not been without challenges (Chamberlain *et al.*, 2006; Alhammad and Moreno, 2022; Cruz *et al.*, 2021; Dobrigkeit and Paula, 2019) as both agile and scrum prioritise quick and fast execution over deeper investigation, which is often needed in design explorations, or are in contrast to ethics toolkits that expect project teams to anticipate far into the future, as explained in the next section.

Where to start?

One idea of ethics tools, interventions and processes seems to be that ethics need to be considered from the start of a project, in order to guide the overall direction of the future tasks. On paper this makes sense, as in the ideal implementation of process tools such as scrum and agile, there are project kick offs, and clear points of where a task and a project is considered "done". In practice however, the observed teams never actually did kick off projects with a clean start, and in fact, it is difficult to pinpoint the exact "beginning" of a project. Is the beginning when teams start to develop? Is it when teams start to research and investigate requirements? Is it when potential clients or collaborators pitch the idea of a project?

In most cases, the team was presented with project ideas that other parts of the city wanted them to develop for them. In theses cases, the project has already "begun", at least in the heads of others, who have done work to identify a problem, ideate around potential solutions and define potential requirements, even if those are not translated yet to technical requirements. From here, one individual of the team takes up the conversation with the potential collaborators to understand what they are after, potentially already starts sketching some code, and reports back to the team, who then starts to plan to incorporate the project into their backlog. At that point, it might still take another two or three steps between management and planing and the tech team actually getting started on development, and somewhere in between there, ethics tools and sessions are supposed to happen to give the team space to reflect and deliberate about what could potentially go wrong and how they would address that. However by that time quite a few important decisions are already made and the team is presented with anything but a clean slate to start with.

Just as with other project management tools, there exists an ideal, "on paper" version of such ethics interventions, and the practiced variant, which tries to account for the non-perfect starting conditions teams encounter. The team reacted to such discrepancies with feelings of guilt and inadequacy, in particular when they felt they did not give the conversation around ethical considerations enough time early enough. Tweaks and changes in other work process where experienced as annoying, but only as relevant for their own way of working. Differences between "ideal" engagement with the ethics toolkits and actual engagement came with feelings of discomfort for not acting responsible enough, surfacing the negative consequences of increasing awareness around ethical considerations and ethical sensitivity.

Where to stop?

Projects in management processes come with a ideal defined conclusion, at which a project is supposed to be considered fulfilled. This is defined by the scope of project, logistical demands etc. These aspects create a direct stream of production, into which ethical considerations need to be fitted, constructing a downstream character for ethical action. First, decisions about what the right action are being made, then they are planned, executed and concluded. But as the teams encounter, that is not really the nature of ethical dilemmas, and any decision or action taken feels like it is falling short of what the actual "right thing" to do would be.

However, ethical dilemmas do not actually have a scope. That concept does not really make sense to apply to ethical deliberation - but in order to make ethics tools fit with the established ways of working, teams have to decide where they define the cut off points, the "good enough", for any ethical challenge they encounter, and where they decide that everything beyond that is not in their responsibility, capability, mandate etc.

The team copes with this disconnect between ideal practice and IRL practice by accepting non-perfect things, arranging themselves with "good enough" and "making do" with what the process allows them. All of this squeezing of ethical tools into existing working structures however also creates clear openings for criticism and frustration, of themselves, each other and resulting works. None which would never be able to live up to the ideal vision painted by ethical interventions and tools, thereby adding to the moral stress through additional potential for scrutiny and potential failure to live up to the high bar set by manifestos and toolkits in regards to responsibilities the teams are asked to live up to.

10.5.2 Internal relational security

This leads us to the important question of how practitioners respond and cope with their experiences of frustrations and falling short. As part of my visit I also collaborated with a second team, also working on developing prototypes for AI driven services for citizens. Within this team, tensions and conflicts were running high, and in contrast to the first team observed, only few of the team members felt invested in the ethical dimensions of their work. When questioned about the topic, the stated that "we are always working human-centred, so we are already ethical" and "we don't work with sensitive data, so we don't need to consider ethics".

The depth of ethical questioning of their work that the first team displayed was not accessible to the second team, and also not considered necessary. In comparison to the first team, the second team had much lower social cohesion, and struggled with a shared agreement over everyone's roles and responsibilities. In such a social space, in which interactions are fraught and the team members do not feel they can trust on mutual support, it appears logical that openness to ethical debates, which might lead to further discomforts and vulnerabilities, declines (similarly discussed by Popova *et al.*, 2022). While the first team responded to moral stress with mutual support, such as assurance that they appreciated each others questions and concerns, or expressing that others did a good job in regards to moral dilemmas, the second team could not create a shared social space in which doubts, concerns and potential fallibility were possible to show.

Operating with discomfort means to take action despite not knowing exactly what the right thing is to do, and feeling accepted for what such uncertainty might bring to the team - effectively a collaborative moving between certainty and uncertainty (Shklovski and Némethy, 2023). This includes being secure enough in delivering ethically "imperfect" work to external audience as well as to colleagues who share ones workdays. This is not possible when team members don't support each other, don't accept each others roles and contributions and don't recognise the good will and good intentions in each other. In other words, the second team displayed dynamics of shutting down and retreating from even engaging with dilemmas in order to protect themselves (Morley *et al.*, 2019).

10.6 Implications

From the described observations, it becomes clear that interventions such as toolkits, workshops and other initiatives to encourage ethical conduct encounter challenges similar to any new practice and change of doing things - such as mismatch with existing cultures, work structures and management processes - while also differing significantly in particular aspects, one of which is the emotional experience of developing increased moral sensitivity and ethical awareness. New work practices always lead to a questioning of old structures and require emotional labor to account for adaption, and ethics interventions add on top that old structures are being questioned on moral grounds and in contrast to what is generally prized with business practice. Even where toolkits and practices for ethical reflection are designed with logistical necessities in mind and technologists are emotionally invested and open to taking on responsibility, interventions aiming to make technologists ethically aware are missing important parts of the emotional experience they create. Such emotional labor produces moral stress, manifesting as vulnerability, discomfort, and other difficult emotions.

However, ethical toolkits often treat their users as disembodied, impartial actors who are expected to engage in reflection, identity work, and required action without emotional reactions to shifts in perspective and awareness of positionality and responsibility, and hence, without experiencing moral stress. As moral stress as the byproduct of ethical awareness and desire for change remains unacknowledged, the fact that the engagement with ethical dilemmas is hard and difficult is interpreted by the team as signs that they are not doing it right - adding additional discomfort and feelings of failing. I argue that for ethics initiatives to be successful and sustainable, they must account for this inevitable discomfort and moral stress. In order to avoid additional negative experience because of a mismatch of expectation and actual experience of ethics efforts, teams need support not only in how to "do" ethics but also in navigating the affective experiences they encounter, reassuring them that uncomfortable emotions are an inherent part of ethical engagement, not a sign of misconduct.

Finally, ethics interventions are frequently confined to technical teams, without expectations for broader organizational practice changes. This contributes to the aspects that are challenging about changing practices and perspectives on work - singular teams can not be the only ones doing the "ethical thing", they need a supportive organisation that adjusts its cultures and assumptions in particular about technical products and innovation processes. While organizations might support their technical teams with extra time and resources for integrating new ethics-related tools and processes, few address the need for wider organizational change (Drage *et al.*, 2024). Simply making technicians "ethical" does not solve the problems with technologies that such organisations are encountering, such as larger socio-technical issues such as privacy, equity and bias, because they are not purely technical problems (Gillespie, 2020), and accordingly the cultures that build cultures need to change as well. Teams that embrace moral awareness and try to influence organizational decisions often face barriers like managerial hierarchies and internal politics, which are commonly not understood to be part of what contributes to ethical issues expressed in technology build by such organisations. This is then were what is considered "ethics" in organisational practice encounters what might be considered office politics, in fact highlighting that separation between the two is fluid.

Conclusion Part 2

11

Both paper 3 and paper 4 highlight that creating interventions for ethical engagement in practice is hardly straight forward, and many approaches are underestimating the far reaching complexity that is uncovered when practitioners start to challenge established notions of how things are done or how certain values are being addressed. As paper 3 demonstrates, values such as trust or accountability are non static, emergent qualities that cannot be with certainty programmed into technologies, because they cannot be with certainty defined independent of context and situation. Interventions aiming to create more ethical awareness therefore require attention to the fluid and shifting tensions amongst values. In addition, they need to highlight the relational parameters that contribute to this shifting of value interpretations and the experiences that result from value judgements. However, methods that engage with this do not end up delivering clear solutions, but fuzzy feelings of uncertainty, as paper 4 shows. Exposing technologists to interventions aiming for strengthening moral awareness leads to moral stress, even when methods are not as speculative as the one presented in paper 3.

In addition, the relational network within which technologists are positioned does not only define stressors such as logistics and resources, but also identity and perception of personal moral performance. In order to deepen the understanding of how ethics initiatives can address the sociopolitical dimensions of organizational practices, practitioners need to be recognised as emotional beings with affective needs when dealing with responsibilities, ethical tensions, and uncertain consequences, in particular within organisational structures which are likely no designed to support a breach with established cultures prising narratives of innovation and progress.

This leads us back to imaginaries - expecting technologists to reflect upon the ethical dimensions of their work and the sociopolitical impacts of it without changing the imaginaries of such technologies and what the role and mandates of technologists are, leaves them hanging.

Part III

Ethics as situated practice

Discussion 12

The increasing popularity of the term "ethical design" insinuates that there is a duality in design practice where you can have design and ethical design. In one interpretation of that duality, you can have ethically neutral design and ethically better design, OR ethical vs unethical design. In either case, this duality does not really hold up. ALL design is ethical design in the way that it makes ethical judgements, and designers make ethical moves, and designs and technology can be ethical or unethical in various ways.

As plenty of prior scholarship has already expressed, technologies have values inherently embedded in them (Winner, 2007), and design practices have values embedded in them just as much (Parvin and Pollock, 2020), so that observation is not new by any means. Ethics is not a static component that can be designed into technology or a problem to be solved through technical measures (Cunningham *et al.*, 2023), it is an ongoing process in which automated decision systems engage (Dignum, 2022). In the case of AI, automated decision systems already engage in a process of ethics when they categorise and evaluate people based on the parameters we give them (Bowker and Star, 1999) - to them, these parameters present ground truth (Zając *et al.*, 2023), and are the only existing dimensions on which they can make such value judgements. In this sense designers and technologists always engage in ethics by design - to design without a process of making value judgments is impossible (Löwgren and Stolterman, 2007). As designers, we are always engaging in ethical reflections and processes (Dourish *et al.*, 2004), by deciding at every product decision what is more in line with the values currently deemed important.

Hence, asking how to design technology ethically is the wrong question to ask. Both aspects of the question are intertwined, and both will happen inevitably. What we should ask is what kind of ethics are we using when we design, and how do our ways of designing engage them? The proliferation of "Ethical Design" can then be understood as design that moves ethical considerations into focus, and intentively pays attention to which moral quandaries are being considered and which are being ignored, which social and political aspects are being cared for and which are not (as pointed out by Mol in her discussion of the politics of care specifically (2008)). In other words, rather than ethics becoming an object of design, or a way of designing, it becomes a point of focus, which is highlighted of particular importance to pay attention to.

12.1 Ethics in Practice

Talking about ethics in practice denotes that there exist ethics outside of practice. Principles, values and guidelines are the outcomes of ethical deliberation with the intention of creating normative guidelines that ensure a certain morally aligned conduct. They exist outside of practice, and they have to, in order to keep their generality. This is normative ethical inquiry about where we want to go. Ethics in practice then puts such representations into relation with the given context of an ethical decision making process.

The important aspect here is to recognise what the impact is of calling out *ethics in practice*. As stated above, practice can't exist without ethics, therefore considering ethics as a practice does not make sense, as all practice includes ethics, and ethics can not be practiced as its own mode of doing - which would imply that there is a way of practicing that is free of ethics, and ethically neutral, which is not possible.

In other words, just as ethics is already part of designing it is also always part of practice, and it is a constant part of all practices that have been covered in the presented papers (and much of the discussed related work). My work has raised this part to the surface for further inspection of the dynamics and behaviours around them. Strategies that argue for concepts such as ethics-by-infrastructure (Morley *et al.*, 2021) or ethical-by-education (Vanhée and Borit, 2022) fall flat with this frame of thinking about ethics. My research has therefore tried to look into where ethics in the form of "paying attention to" is happening, in maybe unintentional ways, and where these ethics express themselves in unquestioned manners. As such, my research does not present a way to integrate ethics into practice or present ethics as an approach, given that those kind of formulations don't really make sense given the inherent nature of ethics in any design that is already happening. Instead, my research has worked out where and how ethics as a process of morally motivated decision making are already happening in practice, and how we can raise those parts to attention to engage deliberately with them.

12.1.1 Ethics as Situated Practice

Building on this understanding of design, ethics and practice as inherently intertwined, my work has been to identify and highlight the dynamics and behaviors associated with them. Based on this examination, ethics in practice emerges as always situated (or, in Suchmans terms, as situated action (1987)): **contextual, relational, particular and ongoing.**

Relational:
people figure things out socially, amongst each other. While values and principles can be held by an individual, practice will always affect others, and hence manifests rela-

tionally, as in within the relations amongst people - in the form of communications, negotiations and potentially conflict.

Contextual:
decisions are shaped by the circumstances within they are made, and many structural aspects influence which decisions are possible to be made and acted upon, such as social, political, economical and historical structures expressed through organisational and physical infrastructures.

Particular:
Where values and principles are applied to context, they move from universal ideal to particular manifestation of that ideal - and take on a particular form dictated by the relational and contextual circumstances of a given context. It hence becomes unique, and this particular instance of ethical decision making wont come about again - even though precedents can make powerful tools to ensure likeliness of decisions in similar relational and contextual circumstances.

Ongoing:
Ethics is not a problem to be solved, it is a process that needs to be constantly engaged in. What is required are iterative interventions, which tweak and course correct as we carefully test our way forward. This should not be equated with any approach such as fail fast and break things - being intentional about trade offs and consequences opens a new window for negotiating which direction we actually want to go.

Recognising ethics as part of practice allows us to change practice in ways that make more space for reflection, questioning and intentionally re-directing efforts and design decisions. Again, I am not the first to argue this (Shklovski and Némethy, 2023; Frauenberger *et al.*, 2017; Wong, 2021), though it is a point that deserves more prominence. It is also important for the part that comes next – considering where to go from this insight, and what it means do the interventions we might design to provide actionable steps forward.

12.2 No framework, but guide posts: Imaginaries, Tensions and Emotional Experiences

With this insight of ethics being understood as a particular part of situated practice, with its own texture and being shaped by similar powers, it is relevant to ask what it means for attending to this part. My research shows that approaches that don't recognise the textural difference between ethics in practice and outside of practice, between normative moral inquiry and working on translating such inquiry into design, will leave practitioners without productive ways forward (Mittelstadt, 2019). And methods and interventions that don't acknowledge the inherent presence of ethics in practice result in confusion and frustrations, which is very much counter productive to the goal of getting more practitioners invested in changing their practices. I suggest that what is needed is not another set of methods or toolkits, nor a framework or process, but rather guide posts for orientation for the transition between these different layers of practice and as pointers to where formative influences are active for ethical decision making.

Based on my research I suggest three such guide posts: Imaginaries, Tensions and Affective Experiences. These suggested guide posts are not supposed to guide a direction, or have normative impact towards which direction ethical decision making is supposed to go. Rather, they are meant to highlight where levers for change can be positioned and might find more productive traction than many current efforts.

12.2.1 The Generalisable: Imaginaries

Imaginaries provide strong framing devices for the interpretation of values and principles. When ethics outside of practice - principles, values - need to be brought into applied parts, those imaginaries provide a bridge to visualise such general expressions of ethics in the shape of a contextualised, desirable future. As Irani (2019) states, sociotechnical imaginaries of innovation for example shape the way people think about development, progress, and economic opportunities. These imaginaries influence individual aspirations, often promoting a particular vision of modernity and progress that currently aligns with neoliberal values, marginalizing local knowledge and practices, and reinforcing existing power structures and social inequalities (Joyce *et al.*, 2021). Engaging critically with these existing imaginaries highlights the importance of understanding the cultural and ideological dimensions of narratives for example of innovation or of AI, and the need to question whose interests are being served by dominant narratives and practices. Such imaginaries can be seductive and distracting, and so to create a productive space to rethink and reflect on ethical implications with the intention of shaping new futures, existing imaginaries need to be properly identified in order to be challenged.

12.2.2 The In Between: Tensions

Creating spaces to challenge such imaginaries allows to bring out the **tensions** that arise from contextualising values, and the trade offs that emerge in relevant ethical dilemmas. When bridging the space between a desirable future defined by certain values and principles and the situated context designers find themselves in, the focus needs to shift from values to value tensions. Tensions highlight where social reality and envisioned futures are conflicting, and where ideals and "good enough" approaches need to meet. Engaging with such critiques highlights tensions on a broader scale, but considering ethics in practice also demands project-focused investigations of situational and relational tensions, which shift from context to context, in particular when considering use cases of AI systems (which I want to also call out as politics in other words). Working through such tensions is an important step in creating moral awareness and ethical sensitivity, which are generally required in order to strengthen intentional ethical decision making.

12.2.3 The Particular: Affective Experiences in Practice

Finally, engaging such tensions outside of safe and speculative spaces creates **moral stress** and uncomfortable feelings, as any resulting changes in mind carried beyond the spaces of ethical interventions rub up against established norms and practices. As stated before, research on bodily sensations and emotional experiences (Höök, 2018) as relevant inputs for designing technologies are starting to connect such insights to ethical engagement as well (Garrett *et al.*, 2023; Popova *et al.*, 2022). I want to add to such approaches by centering socio-political influences of relations to the consideration of affective experiences. Social and political costs of ethical decision making are starting to be discussed (Widder and Nafus, 2023; Popova *et al.*, 2023) and this highlights the importance of considering ethics initiatives not only as descriptive but also as experienced elements in practice. As discussed in part 2, ethics interventions and materials put moral expectations on technical practitioners, which many of are being confronted with recently. While this might be empowering for some, it also feels daunting for others.

Creating awareness of the relational and emotional costs of demanding changes in projects based on ethical concerns directs us towards research focused on helping designers develop two crucial skills. First, the ability for ethical reflection, and second, the ability to cope with the emotional impact of such reflections. This support enables designers to advocate for their decisions and concerns effectively, care for themselves and others, and recognize when intersocial dynamics are challenging due to conflicts between ethical concerns and existing incentive structures or sociotechnical imaginaries. Furthermore, it helps them address these conflicts responsibly by learning to regulate emotions and affect around ethical engagements and decision-making.

Disentangling the different layers that I reviewed in the findings, allows to pinpoint specific points for reflection and intervention about how ethics unfolds in practice. Separating out these layers also allows to identify how they influence each other, and why it is important to consider them together, rather than individually. For example, efforts to adjust practices to be in line with the outcomes of ethical inquiry through reflection and challenging imaginaries in turn challenges established imaginaries of practice and issues of culture, which ties back into the importance of recognising the formative impact of imaginaries on everyday practice. Efforts that target one layer - engaging imaginaries, methods for reflection and evaluating, and processes for operationalisation, - will fall short otherwise.

Conclusion 13

The ethical design of technology has been an ongoing quest in HCI, but has received new surges of interest in the last years. Especially given a new hype around AI releases, academics, governance and activists share concerns that current efforts to "ethically" align the design and implementation of data driven technologies, and AI in particular, is lacking the necessary tools and understandings to effectively address deep seated social, ecological and economic harms. Efforts and interventions such as guidelines, tool kits, and regulations seem to not deliver the desired impact and many challenges around undesired effects of technology on our socio-political systems remain. The question that this thesis then set out to answer is how can we address the question of ethics in the design of AI technology otherwise?

In response to this question, and a set of sub questions, I have reported on four studies that investigate how ethics are integrated in the practice of designing technologies through different lenses: sociotechnical imaginaries, value tensions and affective experiences. Taken together, these studies find that current efforts of addressing ethics in practice fall short of considering the breath of factors that affect how practitioners engage with ethical decision making. My work contributes and extends the work done by fellow scholars of ethics-in-practice, as well as to the broader field of design and AI ethics.

By combining the insights from all chapters in this thesis, I suggest a re-configuration of what is considered within this frame of ethics in practice, in particular in regards to concerns around fluid and relational ethical dilemmas as surfaced by AI. I argue that in order to understand how ethical engagement in technical practice of AI is shaped, we need to not only consider what we deem morally desirable, but also which visions of AI driven futures are being applied to argue for and against other moral values, in the form of sociotechnical imaginaries, and what kind of affective experience the embodied engagement with ethical tensions creates for practitioners. I call these additional themes guide posts, in the sense that they can indicate points of importance for consideration to productively engage ethical dilemmas settled between the generalisable and the particular.

The implications of this suggestion are the following: ethics in practice needs to be considered very differently than ethics as normative inquiry, and put a lot more demands on both practitioners and organisations aiming to enable ethically aware technical practice in the field of AI development. In fact, considering the three suggested guide posts will not make ethical engagement easier or simpler, but more complex. Bringing this complexity to the surface also highlights the political dimensions of this broader engagement, thereby making the ongoing conversation potentially even more fraught.

While this thesis aims to take a quite broad account of different aspects of ethical practice, it is limited by its setting exclusively in Europe, as well as its focus on industry practice. The later does currently not account for other important and influential movements such as increasing regulation within the field of AI governance. Also, this thesis has not taken steps towards providing normative commitment or guidance for practitioners, and therefore is limited to suggestions as to what to pay attention to and care about.

This is where future work is still needed. Further case studies for how to integrate such work spanning imaginaries and affective experiences, thematisizing tensions over alignment etc in practice, are required to provide more actionable approaches on how the suggested guide posts can actually become productive. Following the results of my work, I see the following directions as promising:

A stronger focus on bridging academic research and practice. Lots remains to be done to make any deep insight parsable for practitioners, and not all of that work has to be done by researchers. Design practice requires a new influential turn that shapes best practices to similar extents as design thinking did way back when. These new pathways need to address the current gaps between builders and affected, when experiences with technology are no longer static. And finally, this turn needs to not just discuss the ethical, but actually name the political. Ethics in practice are in the end politics and there is no sense denying it.

At the same time, future work particular in the area of AI ethics might focus on creating a stronger overlap between the ones building the systems and the ones being affected by it. The current strong separation between technical experts being encouraged to think about the affected and the affected being studied in order to understand their lived experiences of AI systems does not lend itself to account for the broader framing of ethics in practice I have suggested in this thesis.

I want to close this conclusion with some personal reflections. Ethics will never be a problem to be solved, and hence, any approach that sets out to find a solution is doomed to fail. I hope however, that my research has shown some potential directions that can put people unfamiliar with the deeper philosophical layers of morality on a more productive, consequential and pragmatic path. To say that it is idealistic is not to say that it is unrealistic - so long as we appreciate that imagining, collaborating and feeling are all necessary parts of this journey.

With this thesis I hope to deliver a well rounded contribution to the quest of figuring out how to deal with the ethics of technology design - one that bridges different layers of ethical decision making, from imagining futures to situated, concrete dilemmas encountered by the technologists that are trying to make a difference. In a way, this thesis is both radically critical of current industry practices, calling out the ways in which systems and structures in which design is practices is preframing what is possible to questions and change, and deeply compassionate to the challenges individuals face when wanting to do better, in which ever way. I believe that it takes both perspectives at the same time, to do the challenge justice, and I hope, that by taking on that duality, the presented insights can actually contribute towards making a difference.

Part IV

Papers

14

Paper 1: Situating Imaginaries of Values in / of / through Design

Title

Situating Imaginaries of Values in / of / through Design

Authors

Sonja Rattay, Marco C. Rozendaal, Irina Shklovski

DOI

https://doi.org/10.21606/drs.2024.803

Venue

DRS 2024 Boston, Biennial Conference Series

Published

June 2024

Abstract

Within the last decade a large corpus of work in HCI as well as the commercial design practice has focused on systematically addressing questions of ethics, values and moral considerations embedded in the design of digital technologies. Recent critiques have highlighted that these efforts fall short in creating actual transformative impact. We use the sociological concept of imaginaries to argue that value and ethics work needs to be considered within the larger context of socially shared visions of a desirable future and outline how existing sociotechnical imaginaries preframe contexts in which value sensitive practices are deployed. We demonstrate that imaginaries provide the language and conceptual framework necessary to address underlying ethical worldviews before ethics-driven design methods and toolkits can be successfully deployed. Finally, we suggest how to engage imaginaries to facilitate a broader shift towards a more politically sensitive approaches in design.

14.1 Introduction

The fast-paced progress of digital technologies has led to growing worries over the ethical dimensions of these technologies and their societal and political impacts. In response, a growing body of work is trying to understand the dynamics that shape the ethical design of technologies (Fiesler *et al.*, 2016; Kahn and Jr, 2007; Kronqvist and Rousi, 2023; Luján Escalante *et al.*, 2022; Shilton, 2018b) and how these dynamics can be supported with tools (Ayling and Chapman, 2022; Gray *et al.*, 2023; Wong *et al.*, 2023), methods (Reijers *et al.*, 2018; Winkler and Spiekermann, 2021; Wong, 2021) and process frameworks (Boess and Jansen, 2022) to steer it into a desirable direction. Values and ethics have been considered within HCI and design research for decades, leading to a large corpus of mature methods and toolkits aimed to support value and ethics sensitive design work (Gray *et al.*, 2023; Rattay *et al.*, 2023; Shilton, 2018b; Stark, 2021; Taylor and Dencik, 2020). We call this collection of practices, interventions and tools aiming to identify and align de-sirable social and cultural values in technology, 'value work'. This work has largely focused on ethical principles and moral values as guiding rails. Recently, the notion of value align-ment emerging from this work has become the main focus of responsible technology design and development. As previous research shows however, values and their enactments shift depending on context and social relations involved (Le Dantec *et al.*, 2009; Manders-Huits, 2011; Winkler and Spiekermann, 2021). Yet it is still unclear how ethical dimensions in design processes are collectively constructed, legitimized and operationalized to conduct such value work. Value work is generally intended to ensure that the consequential impact of technology is positive and leading to a desirable state of future, and scholars have notably argued that imaged futures and cultural values reflect each other (Wong and Jackson, 2015).

We propose imaginaries as a useful lens to explore how designers and stakeholders frame and understand morals, ethics and values in their design practice along with the motivations and interpretations of moral reflection before value work in the design process takes place. A large body of work in STS and CSCW has investigated imaginaries of sociotechnical systems (Ballo, 2015; Dahlgren *et al.*, 2021; Husain *et al.*, 2020; Jasanoff, 2015; Lustig, 2019; Vallès-Peris and Domènech, 2020), hence termed sociotechnical imaginaries. This work calls out the critical aspects in which these sociotechnical imaginaries inform how technology is being perceived within our society and connected to visions of desirable future. Another body of work, particular in the field of speculative design, has engaged with alternative imaginaries of technical futures through different means and their importance for designers to conceive of diversions from the status quo (Baumann *et al.*, 2017; Dolejšová *et al.*, 2021; Dörrenbächer *et al.*, 2021; Light, 2021; Speed *et al.*, 2019). However, much of current value work in design practice is disconnected from the discussions of sociotechnical imaginaries, and speculative design work that engages sociotechnical imaginaries is often disconnected from the applied value work in practice. We argue that designers need to make explicit this connection of sociotechnical imaginaries as pre-framing forces for values, in order to better understand and become critically aware of the productive influences of imagined desirable futures onto the value considerations in their design work.

In this article we provide three perspectives on how sociotechnical imaginaries of technology pre-frame how designers engage with values and ethics in their work: for whom to design, what is designed, and how to design. We illustrate each perspective with examples from the Science and Technology Studies (STS) and the Human-Computer Interaction (HCI) literatures, drawing from the extensive corpus of research showing that sociotechnical imaginaries are socially and politically productive. We argue that value work is being done through imaginaries embedded in the design process of technologies, including how designers must navigate these sociotechnical imaginaries during their work in a situated and practical design process, such as negotiating imaginaries and operating within their limitations. Lockton and colleagues (2019) argue that Research through Design (RtD) should con-tribute to creating change not only by identifying openings in behavior but first and foremost by understanding how people think. For example, rather than designing systems that en-courage people to engage in more energy conscious practices, they argue that we need to address how people think about climate change first. Building on this we propose that before engaging in productive and impactful value work, designers first must consider which imaginaries configure their thinking about values. Being critically aware of existing imaginaries allows designers to guide their engagement with values more intentionally. Our approach enables designers to ask critically why particular futures are deemed desirable and for whom. We suggest reflective questions which can ad-dress the political dimensions of the present imaginaries and their role in design practice.

14.2 Imaginaries

Social Imaginaries, as conceptualized by Taylor (2002), are interpretations of reality and potential futures shared by many people. They describe the mental frameworks that communities use to make sense of the world by providing narratives within which people contextualize their experiences to themselves and to each other. Imaginaries therefore influence how a society structures itself, and coordinates its efforts towards the future, both in framing problematic issues and their solutions. The concept of imaginaries has been broadly investigated in STS and CSCW literature on sociotechnical systems and in design work in the form of sociotechnical imaginaries. Specifically sociotechnical imaginaries lay out the cocreated construction of futures at the intersection of different assemblages of science, technology, and society (Jasanoff, 2015). In this paper, we follow Ballo's (2015) usage of sociotechnical imaginaries as "collective visions of desirable and feasible (technoscientific) futures". The concept has been applied to refer to collective visions of technical futures at different scales, from smaller groups to communities and large organization, and different fields of context from governance to healthcare, mobility, block chain and living spaces (Ballo, 2015; E.S. Ruppert, 2018; Husain *et al.*, 2020; Jasanoff and Kim, 2009; Miller, 2020; Nardi and Kow, 2010; Schwennesen, 2019).

These sociotechnical imaginaries cast explicit expectations and roles which the designed technology is then supposed to fulfill (Lockton *et al.*, 2019; Skjølsvold and Lindkvist, 2015). Sociotechnical imaginaries are hence inherently political: they shine light on the overlapping of social expectations from technology and the social structures about to be established through them.

14.3 Imaginaries as Value Work in Design

Designing is a world making activity (Nelson and Stolterman, 2014), aimed at intentional change with a particular outcome. Design fundamentally involves imagination, and many design materials represent imaginative responses to design questions (Blythe, 2017). These responses in terms of which anticipatory changes are to be made define which aspects of the existing world ought to be changed or left alone, and are hence inherently political (Löwgren and Stolterman, 2007). Imaginaries establish mental boundaries for such imagination by engraining aspirational and normative dimensions in the shared perception of futures (Markham, 2021). Imaginaries contain visions of how life ought or not ought to be lived as well as what version of future is deemed desirable. In order to achieve these versions of futures, they express a shared understanding of what kind of (design)actions are legitimized and thus encode a shared interpretation of moral deliberation. By guiding and limiting the way people make sense of their sociotechnical environments, imaginaries become mundane frames of world-making and hence unavoidable when thinking about what drives and shapes design motivations.

Imaginaries shape the space of potential solutions by naturalizing a specific collection of dynamics and logics. They draw boundaries around what is considered reasonable, attainable, and desirable. It is within these boundaries that designers manifest artefacts which push towards what is reasonable, attainable and desirable. Through their work, designer's manifest imaginaries of our everyday life. If design is then targeted at bringing intentional change about, imaginaries can be seen as the partly unintentional lenses that are being applied to the surrounding environment within which this activity is situated. STS scholars have shown that socio-technical imaginaries preconfigure the explored design space, thus limiting the extent to which approaches to ethics can be translated in practice. For example, following the project "One laptop per child" (OLPC), Ames (2019) explains how the imaginary of the precocious boy motivates and guides the development of a laptop, which is positioned by its developing team as the inspired solution to poverty in the third world. Similarly, Dahlgren and colleagues (2021) lay out how design imaginaries of the future automated home position the household as an individualised decision-making entity, thereby prioritizing the development of hyper-personalised gadgets over communal technology. In both cases, individuality and autonomy are implicitly prioritized values, sidelining community structures which might facilitate a different orientation towards care as in proper maintenance and repair facilities for the OLPC, or shared living routines that take each other's needs into account for a smart home. Using autonomous driving, Graf and Sonnberger (2020) illustrate how sociotechnical imaginaries contribute to the construction of stakeholders views of future users and publics and argue that the role of science in these imaginaries is used to address questions of responsibility and acceptance of technology in alignment with the interests of the stakeholders rather than with an interest for potential users.

Each of these examples outlines a case in which the present sociotechnical imaginaries drive the interpretation and translation of values desired by the designers and developers into reality in a way that ultimately does not align with the desired futures of the users. In this sense, if we talk about values and ethics in design, we need to also discuss desirable futures, and the kinds of imaginaries that formulate those. Just as competing moral frame-works can lead to different design results (Albrechtslund, 2007) competing imaginaries can lead to different outcomes if not properly examined and identified. Taking departure from three fundamental components of design practice – users and stakeholders, technology, and the design process itself, we consider three angles for thinking about productive imaginaries at play:

1. For Whom: Imaginaries about users and stakeholders that influence the relations be-tween actors affected by the technology to be designed.

2. What: Imaginaries of technology that pre-frame the design space within which designers might confine themselves.

3. How: Imaginaries of the design process itself and the methods used influence how designers perform and apply both.

14.3.1 For Whom: User and Stakeholder Imaginaries

As outlined by Woolgar (1990) and Akrich (1992) designers and engineers configure the user, both present and potential, throughout the entire process of development. By defining which characteristics of users are relevant for the user context of the developed technology, they inscribe and reinforce these characteristics through the design of the technology. While the concept of users has been critiqued within design and HCI as too limiting to capture the fluid relationships people develop with technology (Baumer and Brubaker, 2017), imaginaries of potential future users continue to manifest in the envisioning of technology and shape the resulting design processes. These imaginaries of users contain re-conceptualizations and beliefs about how people act and what they desire in the present situation, as well as how they are anticipated to behave in the future in response to the newly developed technology.

Skjølsvold and Lindkvist (2015) illustrate the impact of user imaginaries during the design process for a personal dashboard for coordinating a smart electricity grid. While the future end users of the to-be improved product are imagined to be invested, skilled and knowledgeable, the existing users of the current product were found to be too inexperienced and lacking the specific technical competencies, to be involved in the design process. Despite the clash of present users and future imagined users, the image of the future end-users continued to be used as persona's to keep the 'imaginary' driving the project intact, rather than adjusting the process or the product to match the existing users. Similarly Graf and Sonnberger (2020) highlight that producers and designers of autonomous driving position current publics as naïve in their doubts and worries about this technology, due to missing education, while presenting future users as rational consumers, who will respond to education on the promises of autonomous driving with rational decisions based on straight forward pros and cons in regards to mobility goals. Both cases illustrate that the projected traits of future users sustain, even if they do not correspond with the actual ones, and the existence of the improved user imaginary is maintained and performed with effort and commitment (Skjølsvold and Lindkvist, 2015). This effort in maintaining the future user imaginary establishes the expectation of specific user types possessing certain competences, social and economic capital, or cultural attributes. A technology tailored to these factors then configures its users in its own image, or rather its own imaginary.

In addition to the strong commitment to a desirable user, these imaginaries position future users as individualized entities, divorced from their social network and ecosystem. This singling out of the individualistic user is only possible, because of the imagined shift from reluctant present user to educated and skilled future user. Dahlgren and colleagues (2021) trace the characterization of future smart home users to the persona of the techno-hedonist, who prioritizes convenience and efficiency over shared experiences of care and emotional and social labor, that is being performed in shared social spaces such as family homes. Even where personas are considered as a family unit, technology use is framed through fragmented and individual interactions with the systems. Such user imaginaries tend to soften the tensions of social relationships in favor of seamless envisionings of technology. These visions then also do not account for the political dimensions of intersecting identities in the home related to gender, race, culture and socio-economic status.

Following the project "One laptop per child" (OLPC), Ames (2019) lays out how determinist imaginaries of education, childhood and economy come together and crop the design space to the closest social vicinity of the individualized user, sidelining structural poverty as a potential problem space to design in. User problems are viewed as individual painpoints, and are coupled with singular rather than infrastructural solutions, materialized in the new product or service offering. In these sociotechnical imaginaries, users will be better people – more productive, more invested, more motivated – if only they have the right tools to meet the barriers they are facing at the moment – such as a well-designed electricity dashboard, a trustworthy autonomous car, a supportive and efficient smart home or an easily usable laptop, irrespective of other social conditions shaping their lives. Through the unlocking of the new experience, the imagined users are expected to themselves adopt characteristics and traits that the designers invested in these projects deem desirable: they turn into motivated, skilled and educated users, interested in utilizing the provided products and services to the fullest extent , with the desire to live efficient, convenient and optimized lives.

The postulated values of these projects – freedom of mobility, autonomy and curiosity to learn independently, living an efficient and optimized live – sideline other potential futures, such as futures of mobility in which freedom of mobility and flexibility might mean not taking a car or sharing traffic infrastructure with other commuters instead of autonomous cars as agents; a future of education which prioritizes the opportunity of communal learning, through mutual instruction with care and respect for each other's needs; and a living together in which energy needs are considered within broader social needs and not only cost efficiency. While these values are not inherently moral, they are social and cultural, and be-come normative through the applied sociotechnical user imaginaries, which contextualize the acceptable roles in which future users are being cast. They assign specific blueprints for legitimized behavior and desirable expectations toward other actors, often along the lines of socially established and accepted identities and roles (Oudshoorn *et al.*, 2004).

14.3.2 What: Imaginaries of Design Solutions

The previously described user imaginaries of the tech-enabled and individualized user is closely intertwined with the notion of technical determinism. Many designers and social scientists have examined the prevailing deterministic tendencies engrained in current sociotechnical imaginaries, and how they entangle the design and use of technology with the expectations of what is desirable, for whom and for which social contexts (Campolo and Crawford, 2020; Dahlgren *et al.*, 2021).

As previously mentioned, Dahlgren and all (2021) discuss this tendency to focus technological solutions on "individual self-motivated decision making entities" in the context of smart home imaginaries. By prioritizing narratives of automated decision making as services which instantly satisfy individual desires, needs and preferences, the resulting digital technologies favor experiences of convenience, control and choice. The future smart homes envisioned in these imaginaries are thus turned into spaces of convenience, control and choice rather than care and community. These characteristics shape what designers will conceive to put into these spaces - technologies that would limit any of these experiential values are deemed undesirable, hence will be sidelined in designerly exploration. At the same time these narratives tie technological systems – smart home technologies such as smart speakers and smart security systems others – into the only way to access that vision of convenience and ease, in contrast to the otherwise social, messy and collaborative environment that the home would be without it.

Ames (2019) ties this preference of technical solutioning to the charisma that certain technologies and related imaginaries hold. This charisma is constructed in contrast to narratives around the unattractiveness of alternative solutions, often in context of the status quo. In the case of OLPC, the narrative which defines the undesirable aspects in this sociotechnical imaginary is the traditional educational system centering on ‘schools’. By setting up the binary between ‘bad’ school system and ‘good’ laptop, the imaginary becomes deterministic in cutting away any middle ground and establishes a powerful and artificially created binary perception of reality, with a dichotomy in which complex social systems are reduced to either/or choices. This technological determinism assumes that technology in itself contains an autonomous logic of progress independent of social and political circumstances and histories, thus delivering intrinsic value to its users simply through the ends it delivers (Feenberg, 2009).

These determinist imaginaries rooted in technological solutionism have been extensively critiqued within academia and social innovation circles, but remain present and common within design spaces and difficult to challenge (Gillespie, 2020; Markham, 2021; Roberge *et al.*, 2020). Particular in the emerging fields of AI, data driven algorithms are being perceived as superior and infallible solutions to a broad array of issues, re-enforcing the solutionist imaginary of technology with a mythical notion, what Campolo and Crawford (2020) term ‘enchanted determinism’. As outcomes of these imaginaries, the resulting technologies force their perception of the world as necessary reality onto their users, recreating the imaginations that first shaped them as present (Alkhatib, 2021). These anticipatory imaginaries of technology streamline the complex social entanglements of touted values such as trustworthiness, safety and reliability into static and flat narratives which fit well with technical implementation. However in favor of such technical implementation they discount the affective relational and social aspects of everyday life in which people do not follow reduced rational logic of convenience cost and safety but complex intermingling of lived experiences, aspirations and interpersonal histories (Pink, 2020). As framing narratives imaginaries provide differing interpretations for the contextualization of tradeoffs between different values in specific situations (Powell, 2018).

The positioning of technology by designers as a unique problem solver overpowers the openness of the design process to the complex entanglement of various social, political, and economic factors coming together as shaping factors of the situation framed as problematic and overrides other imaginations of how these factors could be involved in addressing what is deemed to be a technical design problem. This solutionist perspective constrains the space within which designers explore potential directions by sidelining important other non-technical factors of the existing situation and predetermines results by positioning a technical artefact as a solution, without considering required support networks of social, political, or economic nature as important components to both problem and solution.

14.3.3 How: Imaginaries of Design Doing

By projecting future end-users and technical solutions produced through imaginaries, particular types of relations are being put into place between various actors, separating them into actors “inside” the design process and actors “outside” of it. Designers on the inside of the design process need to share the same imaginary to drive a design project forward. These shared imaginaries create a sense of exclusivity because the person who understands or draws from an imaginary creates a social bubble of overlapping understandings, making certain imaginaries of end-users and technical solutions legitimate, while closing others down. These shared imaginaries shape work practices and favor certain design approaches and methods over others (Ackermann, 2023).

From an instrumentalist point of view, methods as tools assume that there is a world with particular and definite features, which can be captured and reported without distortion. However methods are created for a purpose, and they are created by advocates with interests (Law *et al.*, 2013). Methods should therefore be seen as being shaped by social realities too, and the aspect of what is being done through them cannot be disentangled from the aspect of who created them. These two aspects manifest in industry as a generally accepted frame of how designing ought to be performed and by whom (Ackermann, 2023; 2021). Established processes such as design thinking and the double diamond rely on a phased process of exploration, ideation, and iteration, enhanced with the notion of a skilled designer who is able to craft and curate throughout these phases what is being taken to next step and what is not.

Who is acknowledged as a skilled designer contributing to these negotiations, which aspects are counted as creative work and what qualifies as designerly method however is embedded in sociopolitical environments, which reinforce existing disparities and power imbalances (Costanza-Chock, 2018; Irani, 2019). Business executives can become designers for a day after a design thinking seminar (Ackermann, 2023; 2021), but local activists creating communal initiatives to improve neighborhoods cannot (Costanza-Chock, 2018; DiSalvo, 2022). In these situated negotiations aspects of expertise, social standing and influence are framed through the legitimacy of methodology, through narratives of business innovation championing progress and technical disruption (Iskander, 2018; Wang, 2021).

A design process is therefore always a situated action and a negotiation of the technopolitical aspects of the product imaginary that is being conjured by the contributing members and their shared interpretation of values. Here, the present sociotechnical imaginaries direct the boundaries within which designers deploy their toolbox –

they influence which pain points are problematized (and hence valued as being worthwhile to address), how discussions around feasibility are being performed (and whose judgement about feasibility is deemed relevant), which limitations are being accepted and which ones are negotiated and questioned. These aspects influence the reality that is assumed to exist and is hence manifested in the methods, making methods a performative mode of rehearsing the realities we might want to design for. This means methods live at an intersection of representation, realities, and advocacies. And, as Law and colleagues (2013) state "until we can find ways of rethinking knowledges, realities, and methods together in the same breath, we won't have the tools that we need to understand the work being done by our methods. Neither will we be able to imagine a social that is radically different."

14.4 4. Discussion: Bridging Values and Imaginaries

The three suggested perspectives, addressing who are considered as end-users, what the technical solution should be, and how this comes to form the design process, show the productive influences of imaginaries on different aspects of value work in design. They also showcase that simply centering values in design processes without considering the imaginaries at play and the ethical framework produced by them, is not enough (Manders-Huits, 2011). Current value work in design is often disconnected from the imaginaries in which it is performed, while design work that uses imaginaries in design research is often disconnect-ed from the applied values work in practice. By ignoring instead of challenging the sociotechnical imaginaries that underlie many design projects, we surrender potentially productive alternative futures to futures dictated by the current imaginaries without any moral deliberation (Markham, 2021). As Husain and colleagues (2020) argue, identifying and debating the socio-political imaginations embodied through techno-political systems is necessary, to cluster and analyze the normative trends that are emerging through them, and argue for or against a course direction.

Imaginaries are more than ideas about the unknown, they are remixes of what is already known to the people who imagine, manifested by current experiences, informed through stories in media and gained through social interactions, which are then applied and extrapo-lated onto the imagined future (Markham, 2021). A remix implies that the result can be re-remixed again, and analogies can be changed, reinterpreted, and evolve in their meaning. Following Lockton and colleagues (2019), who argue that experimenting with metaphors and imaginaries during the research phase of the design process can reshape understandings of issues and challenges that are complex and inherently relational in their structure, we consider this remixing process to be the appropriate point to raise questions about the political dimensions of the present imaginaries, before entering into value work conversations. Further research shows that this process of re-imagining can provide a productive funnel for collaborative social organizing by questioning and re-mixing existing designs of technology (Lampinen *et al.*, 2018; Light, 2021; Lindtner *et al.*, 2012), and using these re-mixes to raise tensions between desired values and the imaginaries framing them (Kow and Lustig, 2018). The re-imagining of imaginaries is the active and creative engaging with potentials of how we encounter our world-to-be. In the exploration of the entanglements of our world, this kind of stance allows for locating responsibilities and accountabilities in the making of futures, and, as a mediating activity, in the design of technologies and our relationships with it.

So when working through the imaginaries that frame mindsets when designing, we need to ask of these imaginaries as well: who benefits from them? Who in these imaginaries is considered to possess natural potential, the access to power and the opportunity for growth? Who is allowed to participate in these imaginaries and to contribute to them? Who is allowed to shape them? And how would they change if others were allowed to contribute more? If we apply these questions to the previous examples of smart home imaginaries, they might be formulated as whose convenience is prioritized in case of tensions? How is convenience defined – as the absence of labor? And if so – which labor? Who benefits by casting this domestic labor as inconvenient and difficult? What skills and capabilities are assumed to make interactions with the technology easy and seamless, and who is generally assumed to already have such capabilities? Who is affected and how by designing choices as individual tasks rather than collaborative? If we apply these questions to the project of OLPC these questions might raise who benefits from the strong diversion of casting schools into a bad light, and laptops as medium of the future? How do these ideas get transported into the community that is assumed to absorb this message? What are the aspects of life in the lives of school children that get ignored in favor of the happy tech-futuristic imaginary?

None of these questions inherently touch upon moral values, but on understandings of the world, which in turn affect the interpretations of social and cultural values. But these questions allow for taking into account histories and social contexts to concepts generally present in technical value work, such as justice, bias and fairness (Birhane and Cummins, 2019), highlighting the relational power dynamics embedded in these imaginaries. Working through these specific framings of understanding and articulating the imaginaries that are being used to frame the problem space through mappings, associations and artistic exercises might open ways of thematizing which parts of the given imaginaries might themselves be problematic.

14.5 5. Conclusion

In this text we have reviewed the concept of sociotechnical imaginaries and laid out how they produce effects within the design process on multiple levels; the understanding of end-users and their problems, the kinds of technologies that solve them, and the methods by which designing comes about. We have argued that changing components of the design process on these levels – technology, methods and user relations – in an attempt to center values in the design process is problematic without considering the sociotechnical imagi-naries in which they are situated, because imaginaries have a way of reinforcing themselves independent of the forms through which they are enacted, and this includes values. We also discussed the tension between sociotechnical imaginaries as limiting frames of mind while designing, and their potential as design material to change the circumstances considered when designing as well as developing a critical perspective on the political dimensions of these imaginaries. As technological systems become more fluid and our experiences with them not as product centered but framed by dynamic and evolving interactions, designers need to account for that in their view of the design space to be explored. Imaginaries pro-vide a lens to extend that view from product focused to dynamic ways of living with tech-nology.

15

Paper 2:
AI as The Final Moral Device

Title

AI as The Final Moral Device: Ethics in Industry AI Imaginaries

Authors

Sonja Rattay, Marco C. Rozendaal, Irina Shklovski

Venue

Big Data and Society

Submitted

December 2023

Abstract

Broad adoption of AI systems across many domains has led to an increasing number of professionals who are non-expert in AI considering what functions AI could perform in the future. We report on the findings of a series of workshops in which management, tech, and design professionals envision sustainable AI futures through scenario building, intended to enable better decisions about integrating AI into novel products and services. We apply Borgmann's device paradigm as a lens to show how participants commodify morals within their scenarios to unburden humanity from the labor of ethical engagement and construct AI as the final moral device, serving a greater good from a position of impartial omniscience. We formulate a framework and vocabulary to describe this work of legitimization and the seductive narratives and tropes active in AI imaginaries. We argue that such construction of AI as the final moral device contributes to the depolitization and purification of AI as a superior moral agent.

15.1 Introduction

Excitement and hopes around the evolution of AI technologies have resulted in many professionals exploring opportunities to embed AI into commercial processes and offerings in socially and ecologically sustainable ways. This drives a demand for events, meet-ups and communities focusing on the overlap between design and AI to push conversations about the possibilities technologies might offer beyond purely utility-oriented notions of product design towards the ethical dimensions of AI. The current state of generative AI tools fuels a speculative discourse around potential uses of AI in the realm of hopeful potentials and doomful risks. Imaginaries of AI have been analyzed on the level of national governance and corporate efforts (Bareis and Katzenbach, 2022; Paltieli, 2022), AI experts (Hautala and Ahlqvist, 2022) and the general public (Kapania *et al.*, 2022; Sartori and Bocca, 2022).

Expectations, projections and concerns of non-expert professionals, who are concerned with utilizing rather than pioneering AI technology in a range of different industries, has received less attention. These groups include design professionals such as user experience, user interface, and service designers, design strategists, business developers, transformation managers, and development strategists. Such professionals might work in positions that enable them to carry general ideas and imaginaries from development efforts into product and business strategies, marketing and operation efforts, and research processes.

Forecasting efforts are trying to map a course of action for when imagined AI systems pass into the realm of reality. Particularly designers looking to dive deeper into the consequences of their designs on a broader scale turn towards futuring methods to enrich their practice and guide development. What is concerning about this hype is the projection of capabilities onto a technology by practitioners who are involved in envisioning this technology without properly understanding it. While new AI tools might present ever evolving features and capabilities, the harms and problems remain similar in cause though they increase in scope and scale. This poses questions as to how far practitioners are able to estimate how the technology they use in their vision building is upholding the status quo and reinforcing existing structures and dynamics, under the disguise of the shiny new (Yang *et al.*, 2020).

This paper investigates the imaginaries of AI and ethics that professionals who are non-expert in AI invoke when engaging in futuring activities within the context of AI and sustainability. We use Borgmann's (1984) device paradigm as a lens to formulate an empirically informed model, which provides a vocabulary and framework for how such professionals envision the contribution of imagined AI systems to the overall benefit of a global society. We apply the device paradigm to identify and describe stereotypical tropes of AI imaginaries and trace their seductive appeal, that come to form the idea of AI as the final moral device. Through this move, moral decision making is turned into a fragmented commodity, accessible through AI as the device which relieves humanity from the labor of ethical deliberation and the engagement with messy and complex sociopolitical problems under the guise and promise of a smarter and wiser system capable of moral judgement. As we trace the assumptions that make this imaginary of AI as the final moral device possible, we highlight political implications that such imaginaries hold, such as the general deficiency of humanity to satisfactorily address major global challenges like climate change, economic inequality etc. Finally, we discuss how such imaginaries come to limit engagement with the political layers creating an AI supported vision of a "good life", which ultimately contributes to a de-politisation and purification of technical progress.

15.2 Related Work

15.2.1 AI Futures

The public controversy around the social, political, and environmental impact of data-driven technologies have prompted industry stakeholders to pay attention to the ethics, responsibility, and futures of the general narrative of AI. Governments and high-tech companies release strategy statements that often declare AI technologies first as inevitable, then necessary, and finally in demand of leadership throughout unknown futures (Bareis and Katzenbach, 2022). Similarly, big tech corporations drive an optimistic narrative of broadly beneficial AI by developing AI Ethics principles and programs of "AI for social good", promoting specific visions of scientific and technological progress (Sartori and Bocca, 2022). In such optimistic innovation narratives, literary fiction (Hermann, 2023) and innovation myths are intertwined in the promotion of AI ideas by its practitioners (Cave and Dihal, 2019; Musa Giuliano, 2020). 'Making futures happen' is a growing mantra across companies and organizations keeping up with competitive pressures of technical market developments (Smith and Ashby, 2020). 'Future readiness' has travelled from boardroom level considerations down to design and research departments developing products and services for anticipated future needs and markets (Reeves *et al.*, 2016).

The tech sector in particular is following the directive of the future present, which frames innovation as the solution to making a close future a reality, thereby making such futures seem real and feasible. But these collective imaginative efforts are not only anticipatory and pro-actively preparing for technological progress. Visions of the future – including those (re)produced by design – embody ideologies and, along with norms and priorities embodied and expressed, shape policy planning, market economies and cultural imaginaries (Maze, 2019).

Efforts of collective imagination are also sites for meaning making and structuring of social relationships and expectations, making the future a cultural fact with performative and structural consequences for the present (Oomen *et al.*, 2022).

As such, imaginative work about the future is also a site of the politics of the future. Social imaginaries (Taylor, 2002) provide a projection on the future that has implications and consequences for the presence in terms of which choices must be made and which resources are allocated to attain a particular future. Imaginaries are frames for embedded moral reflections and assumptions, prioritizing certain paths of ethical decision making over others. Jasanoff (2015) specifies imaginaries at the intersection of the social and the technical as sociotechnical imaginaries which include "the myriad ways in which scientific and technological visions enter into the assemblages of materiality, meaning, and morality that constitute robust forms of social life". Sociotechnical imaginaries capture the moral preferences which are manifested within technical systems as a means to bring these visions about.

Futuring, as a process of anticipating and proactively responding to potential technological developments, has grappled with the inclusion of diverse social imaginations, in which gender, class and race (Tran O'Leary *et al.*, 2019) are critically included as formative factors of how futures might unfold. Mainstream futuring techniques within design work are consistently pulled towards claiming "neutral" ground for the futures they develop, sidelining marginalizing factors in favor of futures that center the convenient, luxurious and effortless lives of techno-centric, modern, male- and western-biased norms. As professionals in areas that influence and shape the interpretation, introduction and distribution of potential AI technologies turn towards speculative methods and forecasting, the political impact of these efforts needs to be considered.

15.2.2 The Device Paradigm

Borgmann (1984) introduces us to the device paradigm: a way of life which is characterized by the use of technical devices, that mediate our experience of the world. To illustrate this point, Borgman provides the now classic example of the hearth in comparison to a technological heating system – while the hearth requires labor in the form chopping wood, cleaning the fireplace, maintaining the fire, in order to provide warmth, a technological heating system simply requires turning a knob to the desired temperature. The technical heating system acts as the device which separates the means and ends of the experience of a warm home. Because of the technical heating system, the means of chopping wood and taking care of the hearth are no longer connected to the desired ends, a warm home. In contrast, the hearth acts as a thing in Borgmann's terms. A thing, in opposition to the device, is deeply rooted within its world and its context, connecting means and ends in a direct way. Without chopping wood and taking care of the hearth there won't be any warmth. Because of this direct multilayered experience, the thing is inseparable from our engagement with it. The fireplace, rather than a heating system, is also a point for social connection, for appreciation, for the desired comfort of the warmth, connecting families in a shared concern for the ambience of their home.

Devices are designed to make life easier, more efficient, and more convenient by providing ends in the form of commodities – experiences which are instantaneously and ubiquitously present and safely and easily accessible. In order to be able to provide ends as commodities, devices are required to be “unobtrusive, concealed, dependable and fool proof” (Borgmann, 1984, p.84). A devices’ inner workings are hidden away in favor of friendly, easy and frictionless access to the desired ends, thus making no demands on skills, knowledge or understanding, but providing a warm home with the simple turn of a dial. By providing the desired ends in the form of commodities, devices separate us from the direct experience of the world around us, as they hide the processes of production of the commodity behind their machinery. This removes consumption of the commodity from consideration of broader entanglements, such as questions of resource demands, hidden labor, or accessibility structures. While we can see the quantity of wood necessary to fire a hearth and appreciate the labor of chopping wood if we have to do it ourselves, it is harder to keep track of resource consumption of a heating system. Stoking and maintaining a fire is labor that directly leads to the experience of warmth, while the production, construction and installation of a heating system presupposes labor that is out of our minds when starting a radiator.

By hiding away the means, devices necessarily and purposefully disengage their users from the world by saving labor and effort. This separation causes disengagement and passive consumption. As a result, effortlessness is viewed critically in Borgmann’s device paradigm, as he correlates it closely with disengagement and ultimately dissatisfaction. Tension, friction and effort gain new value in this perspective as they provide an anchoring and emotional investment into social infrastructures and the experience of the actual circumstances of our environment. Things that provide a particular focus to such emotional investment are what Borgmann (1984) terms “focal things”. Focal things provide inviting ways of engaging with them and in turn put demands on us in the form of presence, attention, patience, endurance, skill, and determination. They exist within complex networks of human and environmental relationships which bring together elements such as achievement and enjoyment, individual and community, mind and body, and body and world. As such, focal practices are what makes life meaningful and rich.

Borgmann provides us with a lens to analyze and discuss how the design of technology can be investigated beyond its immediate use case, and builds a contrast to the usability paradigm in discussing whether and how a technology is designed to contribute to what might be considered a good life (Fallman, 2010). We use the device paradigm as a lens to lay out the logic behind argument towards a desired disburdenment of ethical engagement and responsibility within our empirical work. Introducing the device paradigm helps us analyse the appeal of imagined AI futures which we illustrate through our empirical work. Using the vocabulary of the device paradigm, we develop a framework that traces the political purification of both ethics and AI. While the device paradigm is more descriptive in nature, Borgmann’s concept of focal things and practices also allows us to argue for a contrasting approach to ethics.

15.3 Methodology

15.3.1 Setting

This analysis is based on ethnographic observations collected during of a series of five futuring workshops, organized and facilitated by a Swedish innovation hub and business network, over the course of a month. Increasingly popular in industry settings, futuring workshops provide structured means of anticipation and mapping out potential directions the present might take, in order to respond proactively. In this workshop series, participants imagined possible, preferable, and alternative futures through different scenarios across different time horizons, following an exploration of trends and signals – a collection of present developments that might impact the future. The facilitators described the purpose of the workshops to "gather a diverse group of experts and practitioners to discover and co-create insights and lay ground for future innovations to uncover new ways for harnessing artificial intelligence for sustainability." The outcomes were meant to inform future activities of the innovation hub which seeks to support companies in improving their business through the implementation of AI. The facilitating organization does not have any financial interests in the outcomes of this workshop series but hopes to provide thought leadership and guidance to businesses and organizations grappling with the uncertainties of recent developments in the field of AI and machine learning.

15.3.2 Participants

The workshop facilitators invited participants from companies of different sizes from small startups to large multi-national corporations. Participants spanned professional backgrounds including business research, innovation strategy and product and user experience design. In total 50 people participated in the workshop series, though not all participants attended all workshops. The majority of attendants did not have any expert knowledge in machine learning or other types of data driven algorithms but worked in the tech sector and in businesses which deploy machine learning technology or aim to do so soon. Hence all participants came from a specific community at the intersection of design, tech innovation and business, oriented towards technological progress and driven by ideas of scientific innovation.

Participants listed a collection of overlapping motivations for attending the workshops: excitement about the future potential of AI, looking for specific use cases of these potentials such as the mundane, or new forms of creativity, as well as a general worry, skepticism, and anxiety towards the perceived risks and potential harms of

AI. These motivations are in line with generally observed public imaginaries of AI, tightly associated with western ideas of modernity and settled between AI anxiety and the desire to take a leap of faith in technology (Sartori and Bocca, 2022).

15.3.3 Procedure

The first author was invited to join the workshops as a participant in the roles of a professional designer and a researcher of ethics within design and AI. During the workshops the first author documented conversations and discussions through notes and photo documentation. After each workshop the facilitators shared written versions of the scenarios that participants developed during each workshop in response to different prompts. Prompts were given by the facilitators and based on scenarios and themes discussed in previous workshops. During each workshop, participants worked in smaller groups and then presented the developed scenarios to everyone, resulting in group discussions. Overall, a total of 19 scenarios were developed, of varying length and detail, some including visual representations in the form of sketches, others as short as a paragraph. The workshop scenarios were projected onto different time scales, with the earliest futures being projected only 10 years ahead to 2032, and the furthest 30 years into the future, to 2052. Each workshop had a specific framing, to facilitate a "traveling into and around the future and back" as framed by the facilitators. An overview of the workshops and resulting scenarios can be seen in table 15.1.

This scenario from workshop 3 for example describes a desirable future in 2042, based on the prompt of using a physical object, in this case hazard tape, to imagine an interaction with a desirable AI system. It includes statements about projected AI capabilities, describes what the desirable qualities of the system are and how those qualities lead to a positive impact:

> "In 2042, AI is harnessed to dismantle the capitalist growth market and has been optimised for degrowth of production and labour. In this AI-enabled system lives are lived within the planetary boundaries. An all-encompassing AI entity calculates the real cost of innovation and simulates the consequences of new solutions. We are also able to see options for enabling the exchange of actions to maintain balance in the system. AI enables us to see the systems and their interconnections that we are part of. Everything is traceable." *(W3 S2)*

The scenarios cover a wide range of topics: economics, governance, social relations, and different configurations of social systems. Some scenarios focus on the relationship between AI systems and the individual, while others focus on the positioning of AI amongst different parties, such as businesses, social groups, nations. We use extracts from scenarios as well as quotes from discussions throughout the article

	Framing	Year	Scnrs	Prompts
W 1	Setting the Scene and Sensing into Change	2024 - 2050	0	Signal scanning and trend mapping
W 2	Imagining possible and probable Futures	2050	4	How will leading sustainable organizations of the future leverage AI?
W 3	Envisioning desirable futures	2052	7	Use a physical object to imagine an interaction with a desirable AI system.
W 4	Visiting Alternative futures	2042	4	How can we reframe assumptions about AI?
W 5	Co-creating pathways to change	2032	4	If everything went right, where would we be in 10 years from now?

Table 15.1: Overview of the workshop framing, prompts and resulting scenarios

to illustrate our arguments and provide a sense of their content and diversity. The complete list of scenarios can be found in the appendix.

15.3.4 Analysis

Our dataset included written-out scenarios produced by the participants and observational notes from discussions and reflection rounds across the workshops. After the conclusion of the workshop series, the first author conducted inductive thematic coding of the collected materials: quotes from workshop discussions, ethnographic fieldnotes and observations, scenarios resulting from the foresight workshops, etc. All authors compared and reviewed the results accounting for how the framing of provided prompts guided storytelling. For example, the first observed workshop was framed in terms of future organizations, and hence the scenarios had a stronger focus on speculative organizations such as businesses or political organizations and how they might apply AI. The second workshop focused on preferable futures without any focus on organizations, and the resulting scenarios positioned positive expectations of the integration of AI in different areas of life.

In addition, we compared statements from interim check-ins and closing reflections to map changing attitudes, opinions and understandings over the course of the workshops. We paid particular attention to how worries, possibilities and a sense of urgency crystalized throughout the series in how ethics was being mentioned and considered in relation to the role and function of AI. Initial round of coding showed that AI was often described as an overarching entity, separate from the actual software that creates it. A second round of coding focused on the roles, functions and properties that were projected onto this envisioned AI entity, by clustering statements such as “AI supports . . . ”, “AI enables. . . ” or “AI provides. . . ”. A third round of coding investigated how these roles and functions shape the way in which ethics was positioned through these imaginaries, and how the roles and functions of AI were imagined to bring about desirable futures.

This process led to a three-layered analysis of the collected material. The first layer outlined the positioning of ethics and its conceptual understanding through semantic analysis of the term "ethics" and identified how this relation resulted in an isolation of "ethics" from moral reflection and a binary state of ethical / non-ethical technology. This described the challenging relationship between ethics as a process and ethics as a desired good, reflecting a separation of means and ends within ethical deliberation. The second layer then outlined utopian and dystopian futures that followed from this binary, and how these futures expressed moral commitments, implying values and normative judgements of right and wrong, good and bad. Here, the ascribed benefits of technology brought out the desire of unburdenment from the difficult labor of moral decision making through devices. This then lead to a summary of the functions, roles and capabilities that were being projected onto the AI system as a device that enabled and unlocked ethical futures. The final layer showed what such a relating to AI might mean for the design of AI systems.

15.4 Results

Throughout the workshops, we noted both utopian and dystopian imaginaries that emerged in discussions along with desires, worries, expectations, and assumptions highlighted within the scenarios. As the participants were non-experts in AI, but experts in their professions of design, business development and strategy, our analysis paid attention to how participants referred to the concept of AI beyond technology, but as a means to achieve some desired or undesired future. Considering that foresight is a common component in the practice of designing innovation-oriented products, we paid particular attention to when ethical considerations were being treated as moral deliberations versus being framed as organizational infrastructural components to be designed, and especially to how AI was positioned in speculative organizational infrastructures. From the analysis we describe a process of construction as participants envisioned AI as a moral device. We take this process apart in three steps (Fig 15.1), and describe each part of this progression and elaborate on its implications below:

1. The discursive isolation of "ethics" as a binary component within technology,

2. The breakdown of specific ethical considerations into AI driven narratives, which illustrate the projection of moral capabilities onto technical systems along with its potential roles and functionalities

3. The coming together of both previous steps to carry the overarching imaginary of AI as a moral device, and its three pre-requisites.

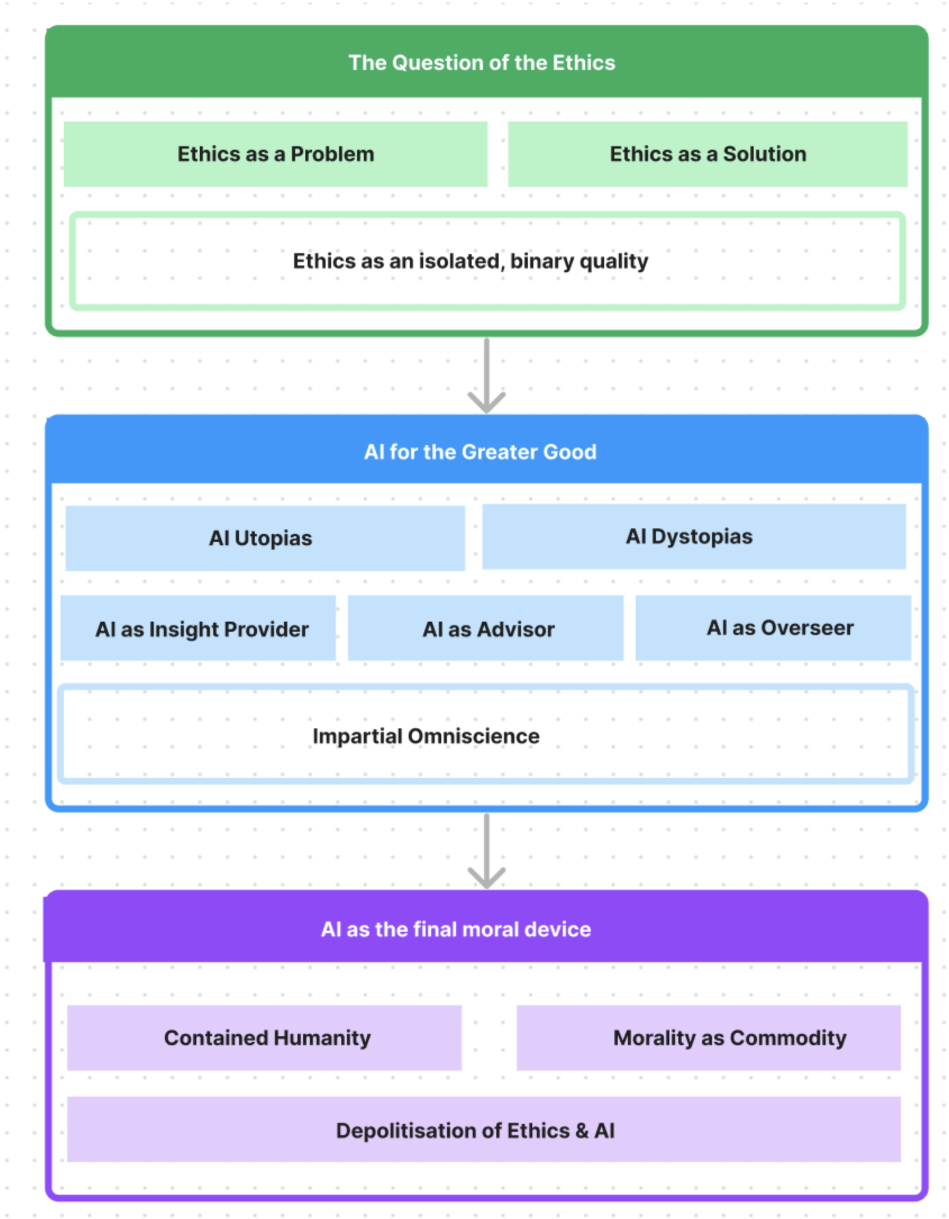


Figure 15.1: Diagram of the argumentative construction of AI as final moral device, illustrated in three main steps

15.4.1 The Question of the Ethics

While the workshop series was framed by the facilitators as developing scenarios using AI in support of sustainability, both facilitators and participants consistently included a range of ethical and political considerations into their speculations as part of envisioned sustainable futures, often blurring the definitions of different terms. While normative moral judgement calls were an important part of the scenario building, participants did not explicitly consider them as 'ethics'. Rather, the term 'ethics' was used as an umbrella term to encapsulate reactionary measures towards imagined circumstances where something could go wrong or be made right, thereby setting up 'ethics' to be considered not as a practice of focal engagement, but as a necessary component for AI as a device. For example, during the summary of a scenario in workshop 2 participants vaguely described their imagined worries as:

> "Along the way, something goes wrong, and we actually need ethics".
> *(W2)*

Yet ethics was also seen as a potentially complicating factor that could be set aside when needed. In workshop 3, participants added as a comment to their envisioned AI product providing on demand advice based on real time collected data:

> "We didn't think about the ethics of it, we took that out to make it easier"
> *(W3)*

This understanding of ethics fits with what complexity theory terms the fallacy of the broken part – the instinct to address a malfunction by finding a singular broken part rather than reviewing the complexities of the underlying system. Such way of thinking underlies many principle or toolkit-based approaches of AI ethics, which consider ethics as a separate field distinct from broader ethical questions (Lauer, 2021). Such attitudes towards ethics can be described as ethical solutionism and ethical determinism (Green, 2021b), in which ethics is either considered to be the necessary fix to technical bugs (we need ethics to fix something going wrong) or as the complication which needs to be addressed separately from the functional layers of the technology (if we don't address the ethics, it will be bad).

Both attitudes follow the dichotomy of means and ends of the device paradigm, describing 'ethics' as an isolated component separate from the overall functionality and roles of AI systems, requiring additional attention to make the imagined system acceptable, rather than an integral activity of imagining a desirable future. 'Ethics' was seen as a response to potential harms and risks or as a layer of the technology that would be deployed as a separate mechanism in addition to functionality. This removed 'ethics' from general conversations around the functionality of AI systems and assumed that the overall motivations and purposes of deploying AI are morally neutral without the need to be questioned as morally justifiable. This lead to an apparent assumption that sociopolitical tensions would not play into the actual design of functionality of AI, and that 'ethics' could be adopted neutrally, without attending to technopolitical entanglements. Participants considered ethics as an isolated instance and as a potential response to sociotechnical problems, separating it from the political dimensions, assuming that any mode of ethical practice would be a suitable response to reduce the risk of unethical outcomes. This separation set up 'ethics' as a commodity, removed from the contextual and spatial relevance (Borgmann, 1993).

Considering ethics as a strongly segregated component allowed for ethics to become a binary rather than a multitude or a pluralistic dimension. In this dualistic understanding, ethics was either an inherent quality of technology (good) or not (bad), drawing clear lines between potential positive and negative futures and eliminating gray zones of contention.

15.4.2 AI for the Greater Good

Whether ethics was viewed as problem to be solved or a solution to be integrated – the result could be discussed as a binary (successful / unsuccessful), which separated interpretations into utopian and dystopian imagined futures. Questions and tonalities that characterized dystopian leaning scenarios touched upon themes such as resource scarcity, loss of lifestyle and life quality, even loss of autonomy and agency. In some dystopian scenarios, deployed AI systems responded to dire circumstances, such as a third world war by governing labor and resources. In other scenarios, AI systems surpassed humanities' ability to govern itself, and were deployed as diplomatic systems between scattered tribes or unified all of humanity in one labor pool in which all resources were optimized in order to meet the circumstances of climate change.

Within utopian scenarios AI was positioned as a singular system calibrated to enable what participants often described as stable balance and harmony – the final desirable mode of an envisioned future. In this imaginary of harmony and balance, AI facilitated a vague notion of the Greater Good, in which needs, desires and interests across humanity and beyond were integrated into an all-encompassing idea of globally balanced wellbeing. Quite a number of scenarios were somewhat ambiguous and were sorted into the utopia or dystopia category by the participants through nomination – the same scenario could be perceived by one participant as a utopia, and by another as a dystopia.

However, no matter whether a scenario was perceived as utopian or dystopian, the general tenor looped back towards a future in which grand socio-political challenges such as climate change, global inequality and exploitative economic models were finally resolved for a future of stable and equitable balance, in which AI facilitated a greater good by evening out power struggles in favor of just and fair living for all:

- Balance between the individual and the collective, in which case AI limited 'selfish' actions, and makes sure that the individual acts in accordance with the good of the greater community.

- Balance between different cultures, in which case AI was used as a political tool to mediate between cultures and groups and take on political tensions. This came in different flavors within the scenarios – either AI created a global unifying narrative and story that made tensions between cultures and communities obsolete, or it enabled a *true* democracy in which conflicts of interest were resolved in the fairest and most harmonious manner.

- Balance between nature and humans in which AI enforces the interests of 'nature' and non-human species in order to force humans to live within the means of the planet and nature.

While the interpretation of scenarios differed depending on whether they were dystopian or utopian, this interpretation only applied to the experience of humans. Given the framing of the workshop series as targeting futures of sustainability, scenarios included varied descriptions of how ecological sustainability would be ensured through AI systems. Throughout all scenarios, utopian and dystopian, nature and the planet would benefit from the intervention of AI, which would balance needs and behaviors always towards even impact – in the utopian scenarios this would also benefit humanity, which would then live in harmony and peace, while in the dystopian scenarios AI would shape the world to the detriment of humanity, which would live in frugal scarcity without agency. In either case, AI would enforce a greater good beyond the needs of humanity, and therefore enact a form of moral behavior which would not be attainable for humanity.

15.4.3 The Roles of AI

Within their iterations on scenarios participants described three main roles which AI could play in facilitating human ethical behavior in service to sustainability. We see these roles as mechanisms of AI as a moral device, the conceptual components that smoothly interoperate without contribution from humans but provide the shiny exterior behind which the turning gears of moral deliberation are hidden. These roles were a direct response to what participants saw as the shortcomings of society and humanity. This is in line with what Borgmann describes as the appeal of technology pulling us towards the device paradigm – the unburdenment from wearisome, difficult, and challenging labor as it is shifted towards technology.

At it's core each role contributes to the idea of overarching balance and harmony as the greater good: (A) insight, (B) advice, and (C) oversight as visible in the lower half of the blue block in Fig 15.1. Fig 15.2 visualizes this same constellation in more detail. Specifically, each role is described through three dimensions that enable their

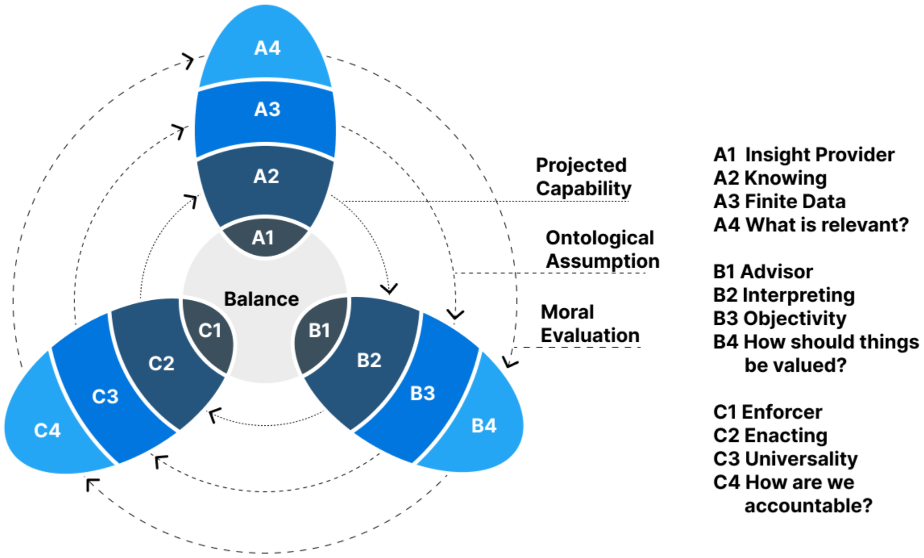


Figure 15.2: Wheel diagram to visualize the layers of the different roles AI takes on within the imaginaries, and how they contribute to the overall narrative.

contribution to the core idea of balance as the greater good: the capabilities that are projected onto the AI system, the ontological assumption that is made about the world that enables that capability, and finally the moral evaluation that this role addresses as a moral agent.

As an Insight Provider (A1), AI systems are helping humanity to "zoom out" from our own perspective towards "the bigger picture" to enable humans to situate themselves within a bigger ecosystem, and see their actions and decisions as connected to broader consequences. Consider the following scenario extracts:

> "AI supports us to gain understanding of nature, regenerate and design ecosystems based on large amounts of data." *(W3 S6)*

> "It (AI) will help us take more parameters into account when making small and big decisions than what human cognition enables us to do today." *(W3 S1)*

In these statements, AI is envisioned to build bridges across nature and humanity through the currency of data, assigning data the main prominence in the projected capability of "understanding" and "knowing" (A2). This envisioning embeds the AI system in every nook and cranny ("small and big decisions"), indiscriminate of the context and situation of the decision in question. The statements imply that humans are constrained in their capabilities to comprehend broader complexity due to our restrictive positionality, unable to take into account the data necessary to gain overarching insight. In contrast, as an all-knowing entity AI possesses complete insight, as it has access to more data than humans could ever process. This idea of AI, accompanied by the idea of real cost, implies that complete knowledge through data is possible, and impact and consequences can be tracked and predicted to their full extent (A3). As a result, it would be possible to map consequences in their entirety, while unintended or uncertain outcomes would be a thing of the past. In this role, AI is used to identify what is of moral importance and consequence based on overarching knowledge, and to appropriately evaluate decisions and actions from a moral perspective (A4). Such imaginaries of AI systems providing simulations, models, and predictive systems to evaluate the true cost of actions speak to a consequentialist view on ethics, basing the moral value of actions on the resulting net harm or net benefit.

As an Advisor (B1), AI systems are imagined as able to gain broader insight and to effectively deploy it for providing advice once data relevant as input for decision making are identified. Consider the following excerpts:

> "[Speculative company] uses AI to calculate emissions, build communities, help to make decisions on processes and materials and provide knowledge of right choices for the organisation and its consumers." *(W2 S3)*

> "AI will be our trusted advisor: it will show alternatives and the consequences of the decisions we are about to make." *(W3 S1)*

Beyond providing insight, AI systems are imagined to facilitate the actual decision process, guiding how the different data are to be interpreted and acted upon. AI systems then would either make decisions independently or support decision-makers in taking the 'right' decision, by providing insights, knowledge, information, and an overview of potential consequences (B2). These 'right' decisions are qualified through expectations of better performance, stressing a belief that ethical action will result in beneficial outcomes within the existing systems. This belief extends the capabilities of the system from 'knowing' more than humans do to also possessing the ability to make decisions and the capacity to interpret the available data in line with moral standards, thus providing appropriate guidance. This of course requires the existence of universally accepted moral standards, thereby assuming that, as a universal and globally connected system, this envisioned AI systems will act on universal agreements of values, norms, and regulations, which will naturally emerge from processing all the available data (B3). Finally, the envisioned AI systems will be neutral and objective, globally connected and taking everything into account, able to provide an unbiased and therefore justified view of the world based on an accepted universal moral standard (B4). This implies that objectivity, and the desired moral standard, can be assumed if only the provided data are sufficient.

As an Enforcer (C), AI systems are imagined to be acting within a stable network of monitoring, automatization, and optimization, to enforce the previously made decisions based on the broader picture insight and to ensure behavior in line with what has been considered "best" (C1). Consider the following excerpts:

> “During the same decade, stricter global agreements and AI-enhanced oversight and reporting of environmental and social impact has made greenwashing impossible.” *(W2 S1)*

> “AI is increasingly used to optimize and automate complex systems across industries leading to new capabilities for systems thinking.” *(W2 S2)*

AI systems will enable optimization, automation, and constant monitoring by tracking resource consumption and coordinating labor efforts. It will also take over boring jobs, to alleviate humans from drudgery. Through tracking and monitoring, such systems will provide full transparency, in which there is no way to hide bad actions or any decision not in line with what the system deems the correct way. However, the AI systems envisioned in the workshop scenarios would allow for a range of behaviors, by allowing negative actions to be compensated with positive ones and weighing selfish acts against acts performed for the benefit of the global community. In this role, AI is used to enforce moral behavior based on the previously prescribed conduct. AI will both enable the implementation of decisions with well defined positive outcomes, as well as ensure that humans are following the appropriate moral conduct.

All three roles presuppose, inform and feed into each other – taking the zoomed out, all knowing perspective enables the system to apply its decision making, which guides the choices of automation, optimization and monitoring. Constant monitoring enables the ongoing and all-reaching data collection enabling knowledge. Such imaginary relies on specific properties which are projected onto the system and the world: universality of values and systems and the assumption of capacity to create data through which the system could achieve being ‘all knowing.’ As a result of this ‘all knowing’, a reliable neutrality and objectivity, an impartial omniscience becomes possible. All three dimensions of balance, between the individual and the collective, between cultures, and between nature and humans, are enabled through the idea of an impartial and omniscient AI system. Such a system is capable of facilitating harmonious living together through the application of oversight and monitoring to limit and prevent harmful human behavior, and to support ‘correct’ decision making that is considered morally ‘good’. Both functions rely on the idea that the AI system can take in and process a totality of data, eventually leading to absolute insight enabling moral superiority. This assumption ties the justification of moral superiority of the envisioned AI system to the idea of objectivity, which is achievable through large amounts of data that only an AI system can process. In this way, the different roles come together to finally enable the god trick, the disembodied view from nowhere (Haraway, 1988), coming together as parts of an automated process to provide morality as a commodity.

Yet at the same time, participants expressed worries about needing to be considerate of ethical and sociopolitical dimensions when developing AI. Towards the end of the workshop series several participants expressed increased concern about run-away scenarios where over-eager use of AI systems to address sociopolitical problems might exacerbate existing sociopolitical and ecological issues, such as considered in this quote from a reflection round:

> “There is so much potential and exciting possibilities [with AI], but ethics has not caught up yet to how fast we develop [AI], so we might develop the wrong things” *(W4)*

While this speaks to the development of increased skepticism and political sensibility throughout the conversations during the workshops, such worries were not articulated beyond diffuse anxieties about potential risks and harms that a well-designed AI system could address, and also remained within the idea of deterministic progress of AI and the separation of ethics from such development.

15.5 AI as the Final Moral Device

Visions of technology as the key towards utopian liberation are not new (Turner, 2010). In the analyzed scenarios, participants echoed the same hopes of AI bringing resolve to humanity. However, the promise of the different functions of AI as an entity of impartial omniscience pushes the utopian notions further by not only considering ethics as means towards satisfactory ends, but also as the end in itself. Morality – in the form of moral understanding, decision making and behavior – is available as an end for AI, thus not only enabling a ‘good life’, meaning a comfortable, safe and happy life (as promised in sociotechnical imaginaries of smart cities, smart homes, or autonomous driving) but also a ‘good life’ in the morally normative sense of every person being able to fulfill their moral duties and to contribute to a society that is harmonious, just and fair. The described roles of AI as insight provider, adviser or overseer illustrate how participants projected their hopes for solutions to the major challenges facing humanity as succinctly summarized by a participant of the workshop as the following wish:

> "We have to figure out how to relinquish control and we have to use AI to solve those problems for us." *(W5)*

Those problems here refer to resource shortages, planetary destruction, and social injustice that participants discussed under the umbrella of sustainability, and were extended towards behaviour deemed necessary to solve them. For AI to be able to attend to this wish, the two previous steps of construction are necessary.

First, moral deliberation is side-stepped by isolating ethics from function and impact as a binary component, which can then be automated. Second, morality is automated as an end in service to a greater good by AI through the capture of reality through data, increase of insight and thus the right to make moral decisions based on impartial omniscience. Ethics can be automated only if it can be isolated and if vast amounts of data and information translate into morally correct decision making. A machine that excels at the latter can perform the former, and justifiably better than humans. Technology can help solve the uncomfortable aspects of life allowing humans to delegate parts of decision-making to artificial agents, extending the scope of their activities towards not only taking on mundane and laborious tasks, but also the difficult, messy and complex tasks of moral deliberation.

In this process of construction, we recognize the relevant components that Borgmann uses to describe a device, and the underlying patterns of the device paradigm: a shift from interpersonal engagement to machinery, and a quantifiable format of morality that turns the latter into a commodity. As a result, AI systems can take the role of the device that will bring the intangible commodified ends of balance and harmony, while putting the means, the actual process of engaging in ethical reflection, deliberation and responding to fellow humans, behind the glossy exterior of a device. The labor of engaging with frustrating and uncomfortable situations in which various needs are accommodated, uncomfortable decisions made, and social structures constantly negotiated, is resolved by the vision of the moral device of the well-meaning and all-knowing AI. Borgmann calls out that "commodified goods, being detached from the burdens of engagement, seem to realise the dreams of magic" echoeing the ease and simplicity which is projected onto living with AI in the scenarios workshop participants created, and highlighted in prior scholarship (Campolo and Crawford, 2020; Musa Giuliano, 2020).

Borgmann's examples of devices focus mostly on tangible objects replacing pre-technological things– a heating system replacing the hearth or fast food replacing a home-cooked meal. In contrast to these tangible items, Borgmann supposes that values and norms have so far remained untouched from the process of devicefying as per their complexity: ethical engagement and moral consideration of how we want to live well together on and with our planet do not have a singular pre-technological counterpart that we are trading for the moral device of omniscient AI. In contrast to a heating system replacing a hearth, values and norms are enacted through actions in every situations of our everyday life, not mirrored in singular objects of importance. However, the separation of values and action is enabled in the imagined scenarios by the separation of insight from values, through the imaginary of data objectivity and moral consideration as means towards the ends of living well together. With the final endpoint of harmonious living in mind, the consideration of AI as a moral device appears logical.

Interestingly, participants imagined the coming together of the different AI roles to enable a return to other pre-technological practices, resulting in pastoral imaginaries of a morally good life, living a simple life in harmony with nature:

> "AI helps us to navigate complexity and rebuild and regenerate different ecosystems. Biodiversity and human diversity thrives. We do not need

> to worry about solving environmental and social power issues and can focus on things that makes us happy, like growing vegetables." *(W3 S7)*

This mirrors other concerns of the one-dimensional interpretation of technological devices through the device paradigm. Verbeek (2002) highlights the ambiguous distinction in Borgmann's philosophy between exertion and meaningfulness. Verbeek lays out that not every technology is necessarily a device but can also act as a thing, depending on the intentions and contexts of use. By making things available more easily and with less exertion, engagement with the ends might also increase rather than diminish. By making commodities more accessible, devices could lead to a further spread and increase of application in people's lives, rather than a general reduction of consideration – applied to our case that might mean that moral agents which provide easier access and closer integration of ethical considerations in people's lives might lead to people considering a broader array of situations and decisions as ethically relevant and deserving of moral inspection. We do not generally disagree with that notion, and also do not state that AI systems would or should act as a device. We argue however that AI systems are currently designed and imagined to be a device. Even where AI systems are envisioned as merely decision-support tools, imaginaries fall back again and again into the seductive pull of the device paradigm in which interactions with AI systems are imagined to unburden humanity and automate away the messy qualities of political life in three ways:

Division of Labor from Result

The envisioned AI systems become the pinnacle of "easily available commodity and a sophisticated and impenetrable machine" (Borgmann, 2000, p. 3) beyond smart desire fulfillment, towards a greater good as the questions of moral deliberation become shaped into a consumption of insight. This leads to a strong division of the necessary labor of building the AI systems from the the consumption of the benefits produced by the AI system. The labour that goes into the construction of AI systems is absent from the imaginaries.

Representation

Abstract ends – such as harmony and balance – become tied to the instruments through which they are attained, such as expressed in this scenario excerpt:

> "Throughout the 2030s and 2040s, the public sector is increasingly using AI leading to a renaissance in democracy as AI enables new ways of citizen participation and direct democracy. Authoritarian regimes are shown to weaken as their political system fails to work collaboratively with AI, using it only as a way to manage humans rather than co-learn."
> *(W2 S3)*

The imagined AI systems here become the gateway to new democratic models, and in turn, regimes who fail to use AI for such democratic work are in consequence failing themselves. Winner (1983) explains the replacement of a need with the projection of the need onto an enabling technology with the example of a car: The need to move around becomes the desire to possess a car, and in return, to drive a car becomes the equivalent to moving around. Similarly, the need for balance and harmony is tied to the instrumented device of the AI system, and in return, the AI system becomes a stand in for what is harmonious and balanced, in contrast to what is perceived as insufficient and flawed, such as humanity.

Contained Humanity

Finally, the imaginary of AI-driven balance and harmony applies AI to optimize the relational, messy qualities of humanity as it equates them with selfish, unreflective, and non-harmonious behavior, contributing to a normative understanding of moral behavior as removed, rational and unemotional. These imaginaries value the way practitioners see AI as structured, rational, and objective, in stark contrast to human qualities. AI enables a world in which humans can remain as they are and do not need to engage in efforts of internal moral growth, as their idiosyncrasies are being externally configured and governed by the AI system with paternalistic exposition:

> "At the same time, AI has learned to write its own code and advanced to the extent that humans are no longer able to understand it. [...] AI governs our societies with the goal of balance with nature and humans inhabit both physical worlds and virtual worlds living fulfilled lives enabled by AI." *(W2 S4)*

In other words, humanity is contained to protect the planet, other species, and parts of humanity itself from its destructive traits. These outcomes were captured both in the utopian scenarios, in which this containment finally enables futures of sustainable equilibrium in which humans also flourish, and in the dystopian scenarios within which AI will make humans either obsolete or turn them into overly optimized cogs in a system. In either case, AI provides the way forward to deal with the general problem of the human condition in favor of sustainable survival of the planet.

15.6 Purifying AI and De-politicising Ethics

The seductive promises of the device paradigm emerge in the future vision of AI as the solution to humanities biggest problems through the imaginaries of non-expert practitioners, who apply the constructed roles as shorthand to wishful thinking. The understanding of technology as a tool to automate uncomfortable and difficult aspects of life is extended from work drudgery to the tricky questions of ethics and politics. The desire for a moral device comes from a place of disillusionment and wishes for a better solution than what people alone have been able to achieve. The readiness to embrace limitations and impacts on humanity shows that there is a willingness to sacrifice comfort and lifestyle to do what is 'right', if that path is clearly laid out by a trustworthy entity. It also stems from the fear of running out of time and not having the right skills and capabilities to solve the challenges we face. Yet such a turn towards automated solutions for sociopolitical challenges results in two counter-productive moves: ontological purification of AI and a de-politicising of ethical tensions.

All of the scenarios in the workshops presupposed that AI systems would be able to gain knowledge that is impossible for humans to gain by other means. As AI systems are imagined to be able to access all available data, thus gaining the coveted certainty of knowing every possibility, this renders the vantage point and values of the receiver of insight, advice or oversight from the AI system unimportant. This is a form of ontological purification, a technological embodiment of the god trick (Haraway, 1988), giving humanity access to 'universal truth' through AI. There is an expected ability of AI systems to "measure and process quantities of data that are too large for human tabulation, too discrete for human perception, and too complex for human cognitive analysis" (Goldenfein, 2019). This extends towards a moral prerogative to uncover some true and hidden state of the world, which promises access to an untainted rational interpretation of reality through which AI becomes an ontological authority that can not be disputed (Goldenfein, 2019). Thus AI systems are imagined to produce a new, pure way of creating knowledge that is as legitimized by technological superiority, allowing to unlink knowledge production and political power (Jasanoff, 2003).

Through such ontological purification the de-politisation of ethics becomes possible, as it presents AI systems as irresistible and uniquely powerful, while at the same time shifting accountability from the people who build systems to the system itself – if the tool is safe and ethical, there is no need to ask how it was built and who built it. Such a view leads to a de-politization of ethics by flattening power imbalances and political tensions into the general idea of contained humanity. Competing needs between humans and other parties such as ecological and more-than-human parties would be resolved through the understanding gained by and through AI systems alongside existing political disparities into equal gain or loss of lifestyle, comfort, opportunities etc. Finally, even in the cases where human labour is seen as an important part of the way forward – the end goal would still be a future in which harmony amongst everyone is maintained without further human effort, making politics unnecessary. This turns the question from not just about how to build ethical / responsible AI, but how to build a machine that promises to take care of ethics and morality for us eternally, turning ethics into the means to maintaining the ends of harmony, and specifically increasing the distance between people and ethics and politics as an ongoing practice.

However, as Borgmann points out: "To speak of technology making promises suggests a substantive view of technology and is misleading" (1984, p.103). Technology can not make promises. People make promises and who makes these promises to whom are political and social justice questions. Claiming that AI systems now will finally resolve existing harms is envisioning a device that solves for us the harms caused by a device focused culture in the first place. As Strong and Higgs (2010) note:

"Engagement with focal things and practices alerts us to the forces opposing them and the flawed use of devices—the irony of technology. We destroy the engagement we enjoyed with them when we try to enrich our lives through consumption." As it so often happens, the goal of the workshop was the discussion about sustainability and ethics, but the outcomes produce imaginaries of AI that would solve the problems of sustainability as ethics for us.

If we focus on the process of ethical engagement, what is put into focus is the means rather than the ends, when the ends are also inherently part of the motivation of ethical guidance. In order to question and contest the ends, the means must be understood. AI as a moral device is a form of deploying technological means to find answers to ethical questions. It contrasts human capabilities as biased, limited, and inherently falling short to the capabilities of the AI system as objective, complete, and hence superior. The imaginary of AI as the final moral device necessarily excludes humans from being able to understand and follow the means. This separation between human means and AI as the means purifies the actions of the AI system from human errors and biases and declares it as consequently more trustworthy, and non-contestable. In this way, AI systems are modeling an utopia of idealized interactions and transfer the labor of fitting into these utopian norms to the outliers and edge-cases - the groups already marginalized and pathologized through systems that do not recognize their lived experiences (Alkhatib, 2021). Without engaging with the political dimensions of underlying social structures, unified approaches to operationalize ethics are bound to reinforce a conservative political status quo (Birhane *et al.*, 2022; Zajko, 2021), where contestation will be ever harder.

15.7 Conclusion

Efforts such as the observed foresight workshops create the spaces for necessary deliberation, exploration and conversation about AI to occur. They demand time investment and interpersonal engagement, facilitating creative thought and critical reflection. Efforts that deliberately bring together influential practitioners at the periphery of expertise in AI technologies contribute to the broadening of circles that appreciate the complexity of the integration of AI systems into our social fabric. Conversations that allow for the creative illustration of the future through scenarios, demos and prototypes can support the development of political sensitivity as contribution to the skillset necessary to establish ethical deliberation as a focal practice beyond checkbox and tool-oriented thinking. However, while such efforts can present an imaginative method of thinking about alternatives, they can also contribute to an existing social status quo by reinforcing interpretations of possible futures and reproducing old paradigms in new clothes.

While the conjured imaginaries sound fantastic and inventive, they also remain within clear structures of political disengagement, thus undermining political imagination in exchange for technical imagination. It appears easier to imagine almighty technologies that solve human struggles, than to imagine humans investing in the work of changing existing socio-political systems through political labour. The mapped roles and functions of AI within imagined AI-supported futures visualise how difficult it is for the participants to break with the dominant deterministic imaginaries of technology shaping our social infrastructures (Markham, 2021). While these imaginaries raise expectations of capacities of technological systems, they reduce the imagined need for political involvement, positioning technological progress as a promise of a comforting fix to structural societal problems (Bareis and Katzenbach, 2022).

Identifying and accounting for who is involved in defining the roles and purposes of AI is a mandatory step towards problematizing such AI imaginaries (Robbins *et al.*, 2016). We must ensure shared ownership and ongoing access to routes of resistance and contestation (Benjamin, 2022; Joyce *et al.*, 2021), particularly in areas strongly dominated by the interests and practical requirements of industry (Gerdes, 2022), which can sideline democratic contestation. Borgmanns' device paradigm provides a powerful lens to recognize what kind of relationship with AI and morality we might have, and where our focus should move to instead – towards the means which are being deployed to construct visions of automated morality as an end, and towards alternative means which prioritize critical and political work on the sociopolitical aspects of AI.

Our framework and vocabulary enable recognising and identifing the seductive narratives and tropes active in AI imaginaries, to highlight where these imaginaries contribute to a disengagement rather than a deeper refocusing on ethical practice. More work is needed to push for ethical consideration to be enacted as an ongoing practice, a continuous process of engaging with the needs of ourselves and others again and again, not with the expectation of reaching a final destination of stable harmony, but as an ongoing investment of our collective social wellbeing. Ethical consideration should be reframed as a focal practice, that might be supported through focal things, as well as technological ones.

16

Paper 3:
Sensing Care Through Design

Title
Sensing Care Through Design: A Speculative Role-play Approach to “Living with” Sensor-supported Care Networks

Authors
Sonja Rattay, Robert Collin, Aditi Surana, Youngsil Lee, Yuxi Liu, Andrea Mauri, Lachlan D Urquhart, John Vines, Cara Wilson, Larissa Pschetz, Marco C. Rozendaal, Irina Shklovski

DOI
https://doi.org/10.1145/3563657.3596066

Venue
DIS ’23, Pittsburgh, PA, USA

Published
July 2023

Abstract
Sensor networks are increasingly commonplace in visions of smart cities and future healthcare systems, promising greater efficiency and increased wellbeing. However, the design of these technologies remains focused on specific users and fragmented by context, overlooking the diversity of needs, wants and values present when technologies, people, and lived realities interact within instrumented spaces. In this paper we present a workshop method – Sensing Care – that can help researchers, interdisciplinary design and development teams, and potentially affected users, to explore what it takes to design for living with sensor technologies that intersect and interact across private and public spaces, through speculative scenarios and role play. Drawing from three deployments of the workshop, we discuss how this approach supports the design of future care-oriented sensor networks, and helps designers understand what it means to live with complex technologies as people traverse diverse contexts.

16.1 Introduction

In the last few years, the city of Amsterdam has deployed over 200 cameras, around 230 air quality sensors, and almost 500 beacons throughout the city. Research projects at AMS, the Institute for Amsterdam Metropolitan Solutions, from "roboboats" to "scan cars", investigate how to integrate mobile sensors within urban spaces, perpetuating the vision of a fully connected, smart and aware city, optimized for the seamless coordination of mobility flows, energy infrastructure and management of city assets. At ACRC, the Advanced Research Center in the city of Edinburgh, researchers are working to embed machine vision sensors into the homes of older adult citizens to enable remote health monitoring and diagnosis, using sensor networks to integrate personal care in medical communication procedures.

While engaging in different settings, the visions associated with smart cities and connected homes paint pictures of urban and private spaces filled with helpful technology systems that assist people and take on the labour of maintaining wellbeing. However, there are growing concerns about the uncertain consequences for people inhabiting spaces instrumented with sensor networks (Cottrill *et al.*, 2020; Dahlgren *et al.*, 2021). Researchers have grappled with privacy concerns (Caire *et al.*, 2016) and value-tensions (Forlano and Mathew, 2014; Friedman and Hendry, 2019) brought about by these technologies – tensions that are likely exacerbated when cities, homes and other private and public spaces are instrumented and networked alongside or with each other, and when people inhabit and traverse these different spaces. Recently, researchers have turned to the notion of entanglement (Frauenberger, 2020) to help understand the complexities of technologies like these, and how they eschew the idea of creating well-defined use cases around singular users.

Frauenberger's (2020) notion of 'entanglement' places humans and things within a system of relations and interdependencies, focusing on the design of configurations of actors, interactions, and environments and understanding the relationships in-between. This re-orientation challenges simplistic conceptions of the user in HCI (Baumer and Brubaker, 2017; Irani *et al.*, 2010; Skirpan *et al.*, 2022) through a shift of perspective from designing for 'use' of sensor networks towards designing for 'living with' sensor networks (McCarthy and Wright, 2004) and the relations between people, things, and environments (Elsden *et al.*, 2017; Frauenberger, 2020). Given the ubiquitous presence of sensor networks, we are long past simply using such technologies in distinct and well-defined situations. When considering an expanded view of what 'living with' sensor networks implies, the level of complexity increases. Consequently, the idea of 'living with' has proved difficult to integrate into technology design practices, not least due to the complexities of technological entanglements within changing contexts of daily living, routine, and extended use. Adopting an entangled and relational way to understanding sensor networks prompts us to consider the situatedness of multiple actors (including the human and nonhuman), how desirable actions and interactions can take place, and how value tensions can emerge from the interdependencies between multiple actors. A broad consideration of values (Friedman and Hendry, 2019), and the tensions that may arise from conflicting ones (Miller *et al.*, 2007), is important to understand the forces that operate within these interdependencies.

In this paper we present a speculative workshop method – Sensing Care – that can help researchers, interdisciplinary design and development teams, and participants, to explore what it takes to design for living with sensor technologies that intersect and interact across private and public spaces. In our work, we drew heavily on prior work on speculative enactments (Elsden *et al.*, 2017), an approach to exploring first-hand interaction and experience with participants through speculative design projects and fictional scenarios. The emphasis speculative enactments place on complex social contexts and the creation of scenarios with consequentiality for participants, means it is placed well to support inquiry into the entanglements with sensor technologies. We developed our adaptation of speculative enactments as part of an ongoing collaboration with two organisations working with sensor networks and associated municipal and in-home services, with a focus on promoting new forms and models of care in later life. In this context, we sought to develop a method that could be used both with external stakeholders and internally within the organizations to break down the common silos between designers, developers, user researchers, data scientists, and others working on different aspects of the same technologies.

Our method brings together speculative narration and role-play as modes of engaged reflection and acting-through scenarios of relational conflicts, combining techniques such as timelines, design cards, and theatre, into one participatory workshop process. Carried out across three deployments with diverse groups of participants, the workshop process provided ways to surface tensions around roles and responsibilities in care sensing, eliciting insights into the contextual nature of trust and the importance of integrating public and private networks of care. In particular, we emphasized the contextual positionality of care and related values such as trust and privacy, bringing a relational dimension to current understandings of these values.

In presenting our method, we contribute an approach to designing for 'living with' sensor networks – and data-driven technologies more generally - in complex contexts that include changing and competing values and priorities such as care contexts. The method provides an accessible means to support citizen participants and diverse research and development teams to reflect on assumptions around use-cases and people related to their projects, enabling design teams to engage purposefully with value tensions and to re-envision the sociotechnical futures they are collectively working towards.

16.2 Related work

16.2.1 Sensor Networks, context, and entanglements

In his seminal text, the Computer for the 21st Century, Mark Weiser (1991) presents a future vision of ubiquitous computing where “the computers themselves vanish into the background”. Weiser envisioned home, urban and workspaces which quietly attend to our needs, anxieties, and inefficiencies, while keeping the interface to a minimum. Cleaving to the vision of the ‘disappearing computer’, context-awareness has become a focus of much ubicomp research and practice (Lopes *et al.*, 2013), in urban navigation (Taylor, 2011), for the mobility of older citizens (Kötteritzsch *et al.*, 2016), and as a means to maintain resident safety in cities (Lee *et al.*, 2007). Similarly, there has been a proliferation of many devices, services, and sensor-systems seen as solutions for the mission of active and healthy aging, automating provision of care through, for example, ambient assisted living (Wynsberghe, 2017) as a way to support aging-in-place.

Despite addressing very different contexts, smart urban and home systems encounter many similar challenges, from being (un)able to correctly interpret context, to being able to react to unforeseen situations (Yao *et al.*, 2019; Yeung *et al.*, 2019). Smart systems can fail disastrously, such as when Tesla’s driverless car autopilot resulted in multiple deaths (Bevilacqua, 2017; Hawkins, 2022) or when biased results from facial recognition software used by law enforcement resulted in erroneous arrests (Johnson, 2022). Although death and incarceration may seem far removed from an errant smart home sensor, the underlying intentions, technologies, and systems used are all a part of contemporary ubiquitous computing, its messiness, and our entanglements with these systems. Issues in these underlying frameworks, algorithms and data acquisition methods (Garvie and Frankle, 2016; Kordzadeh and Ghasemaghaei, 2022; Krumm, 2007) point to an unaddressed need to engage with intertwined social contexts and to be adaptable to contexts yet to be encountered.

Sensor networks as context-aware systems directly challenge the idea of who the user is and how to account for the other “out there” (Burrows *et al.*, 2018; Taylor, 2011). Positioning the user as central has inherently normative implications for how design processes around these systems unfold (Bardzell and Bardzell, 2015). Current instantiations of sensor networked systems attempt to understand contexts and situated interactions through constant data collection, which affects everyone within the vicinity of sensors and poses challenges to notions of consent, trust, and privacy. Yet design typically focuses on direct users, doing little to account for those who do not directly own or control the systems monitoring instrumented spaces, such as guests, passers-by, and other temporary stakeholders (Abdi *et al.*, 2022; Bernd *et al.*, 2022; Yeung *et al.*, 2019). Ideas and visions of the ideal user draw attention away from the cases in which ubiquitous computing encounters situations and contexts where people do not fit with the expected ideals, resulting in overlooked and under-designed scenarios unable to grapple with the relational complexity of shared spaces (Dahlgren *et al.*, 2021; Irani *et al.*, 2010). Accounting for these complexities requires design methodologies which acknowledge and embrace diversity rather than attempting to control and homogenize it (Irani *et al.*, 2010).

There is much agreement that design involves making normative judgments about values, which come to be inscribed in technology (Akrich, 1992; Latour, 1992), highlighting the need for designers to be reflective of, and responsive to the normative impacts of their practices. We focus on 'living with' sensor networks, where value judgments are implicit in how to enable such technologies to support interactions and user experiences that are experienced as 'living well'. Values, such as privacy, trust, or care, manifest differently depending on the situation. The value of privacy, for example, is not abstractly defined or absolute, but instead depends upon maintaining expectations and contextually appropriate interactions between humans, technology, and the flows of information (Nissenbaum, 2004). These contextual shifts are important when considering not just designing for the abstractions of values but designing for values as these are enacted in life, when people live with technologies.

While the notion of entanglement incorporates attention towards value enactments, the question remains how to operationalize these in design practice and what form reflections on embedded values must take when addressing the challenges in the design of sensor networks. To move away from the focus on the user and interactions between the user and their technologies, we need to develop methods that foreground the depth of factors and forces that affect 'living with' technological systems. Designing for entanglements is to "leave user-centered design behind and develop agonistic, participatory speculation methods to design meaningful relations, rather than optimizing user experiences" (Frauenberger, 2020). When designing for large networks of ubiquitous sensors, this notion of entanglement foregrounds the fluid interdependencies that are often presented as unintended consequences to an inherently isolated user-device experience. One way of doing this is by incorporating notions of care, care relations, and care networks into technological development.

16.2.2 Care entanglements

In our work, we are collaborating with partners who are interested in applying networks of sensors to develop solutions for in-home care and municipal services. As such our work connects to a longstanding body of work on the development of ubiquitous computing systems to care services (Key *et al.*, 2021; Light and Akama, 2014). Despite increasing attention to the notion of care in HCI, care remains a difficult term to define. Within ambient sensing environments, and especially where these spaces purport to care, Frauenberger calls for abandoning the notion of user experience entirely and instead designing for relations. In this paper, while care is a service domain of interest, we also ground our work in broader definitions and explorations of care that aligns with the notion of entanglements. We follow the definition of Fisher and Tronto (1990), viewing care “as a species activity that includes everything we do to maintain, continue, and repair our “world” so that we can live in it as well as possible. That world includes our bodies, ourselves, and our environment, all of which we seek to interweave in a complex, life sustaining web.” Such an open-ended conception of care connects with the notion of entanglements, acknowledging the fluidity of relations between actors. The logic of care (Mol, 2008) recognizes that what counts as good or the ideal of a good life is not fixed, but contextual, fluid, and personal.

Care practices have been translated through design experimentation to many contexts, including healthcare and public space (Light, 2021), as well as democratic inquiries in ways to design for relationships between communities and institutions (DiSalvo, 2012; Light and Akama, 2014). To address anxieties around the uncertainties of sociotechnical innovation, institutions are investing in responsible research and innovation efforts, in order to responsively take care of the future (Stilgoe *et al.*, 2013). Rights- and risk-based frameworks attempt to ensure the protection of values of perceived importance, such as privacy, independence, and autonomy, centering people as autonomous and independent individuals who can make informed decisions about technology use given sufficient information. This perspective puts impossible responsibilities on people as they make decisions about technology (Dourish *et al.*, 2004; Seberger *et al.*, 2021). In contrast, the logic of care positions people in relation to each other, entangled in networks of varying needs, relying on each other to make decisions, and embedded in time (Mol, 2008). Yet care is also political. To care means to judge about whom to care for, how, and to which extent (Martin *et al.*, 2015), often leading to value-tensions (Yacchirema *et al.*, 2017) and conflict. This implies that context-aware systems that instrument environments must make space for uncertainty, tensions, and conflict in order “to imagine a world organized to care well” (Tronto, 2010).

16.2.3 Using enactment to design for entanglements

Speculative and critical design work has engaged with notions of care and entanglements through a variety of provocative interventions. Speculative design has a tradition of creating provocations and openings to think about futures and to critique current practices (Auger, 2013), with recent projects engaging audiences by focusing on the limitations and possibilities of everyday life (Chopra *et al.*, 2022; Forlano and Mathew, 2014). Efforts to involve participants as co-creators have led to participatory forms of speculation such as Critical Play (H. Tan *et al.*, 2022) and Participatory Design Fiction (PDFi) (Nägele *et al.*, 2018), bringing attention to messy and conflicting values in real contexts. Some of these approaches have utilized role-play methods to engage diverse stakeholders such as users, providers, decision-makers, and design researchers (Elsden *et al.*, 2017; Vines *et al.*, 2014). Researchers have also explored ‘live’ theatre methods to create shared and embedded learning about the complexity of social values such as care, privacy, and trust through acting within designed possible near futures (Skirpan *et al.*, 2022; Vines *et al.*, 2014). The use of theatre techniques such as improvisation in design workshops offers an entry point to 'unfinished' future situations which can help to create openness for participants (Medler and Magerko, 2010), enabling designers to prototype compelling and dramatic future situations by embedding participants' own values, experiences and concerns (Pschetz *et al.*, 2019; Skirpan *et al.*, 2022; Vines *et al.*, 2014).

Building on this range of speculative design and scenario-based design approaches, Elsden et al. (2017) introduced the notion of speculative enactments, which focus on stage-setting and interventions situated in everyday scenarios, to provide for "grounded, but unscripted improvisation of particular futures". Speculative enactments build on role-play approaches, with each enactment guiding participants through carefully orchestrated experiences of mundane future scenarios. However, speculative enactments lean on leading participants through a critique of a defined scenario, constructed by the design team. Similar to many role-play methods, speculative enactments can be quite constrained (Vines *et al.*, 2014), in an attempt to enable participants to grasp and relate to the experience that have consequentiality (Elsden *et al.*, 2017) for their own lives. Light (2021) distinguishes between more democratic speculative practices and those that are more designer controlled. She calls for approaches that seed rather than lead and may offer glimpses of a range of futures, which allow participants to go in many directions. Inspired by the long tradition of role-play and theatrical techniques in design settings, and the more recent work on speculative enactments, we sought to create a method that can seed (Light, 2021) participatory enquiry into future technologically mediated entanglements around care, situating participants in a future imaginary.

16.3 Research context

The project was initiated as part of a research program between a selection of European universities and industry partners. We worked with two different organisations that develop technical systems based on a range of sensor technologies. Located in Amsterdam, Amsterdam Metropolitan Solutions (AMS) develops sensor networks to help the city become more efficient by supporting optimization of mobility flows, energy infrastructures, and management of city assets. The Advanced Care Research Centre (ACRC) in Edinburgh is working to embed machine vision sensors into the homes of the older citizens with care needs to enable people to live more independently in their own homes. Despite their differences, these organizations rely on similar ideas of instrumenting public or private spaces using emerging sensor network technologies. Due to the inherently invasive nature of embedded sensors, both organizations expressed worries about how future services might address privacy concerns and remain trustworthy. The organizations were seeking future-oriented methods that could enable them to explore notions of trust and privacy through anticipatory scenarios and user-centred explorations.

16.3.1 Industry Partner Interviews

To develop an understanding of our collaborator's needs, we conducted fifteen in-depth interviews over the course of two months with domain experts involved with AMS and ACRC. We selected our interviewees based on their complementary perspectives on sensing networks. At AMS interviewees included citizen scientists, interaction designers, technologists, and city legislators. At ACRC interviewees included healthcare practitioners, nursing studies scholars, social scientists, and technologists. While AMS and ACRC both expressed interest in broad exploration and speculative design work, neither had been able to find productive ways to implement these aspects in their current work practices.

The first five co-authors reviewed and discussed all interviews, conducting iterative, thematic analysis throughout the research process (Braun and Clarke, 2012), thus ensuring that new interviews helped build on the understanding gained from prior interviews. Throughout the interviews, industry partners expressed worries about reinforcing oppressive structures through the implementation of their technology, and how to build awareness and sensitivity towards these dynamics. They were not just worried about how to get the implementation right, but also how to not make other matters worse. We identified four primary concerns: multiple voices and actors, interdependence, situatedness, and performativity, which provided a framework for our design process.

16.3.1.1 Multiple voices and actors

Both organizations aimed to design products and services fitting the imaginaries of a world well-equipped with sensors in the service of citizen or patient care. Many interviewees explained their concerns around the potential unintended consequences of their work in relation to more abstract political and social issues, such as enabling sociability while respecting privacy for people in their homes or providing just-in-time services in a GDPR-compliant way for the citizens. The need to address multiple voices and actors holistically and to attend to interdependence, emerged as we observed the strong fragmentation of research and engineering efforts in both organizations.

16.3.1.2 Interdependence

Technologists and designers involved in producing these systems described a general sense of uncertainty in how to engage productively with systems that can be situated in, and traverse between a variety of use contexts and the value tensions that can arise from it. Very few considered the interconnections between the technologies they were researching and building, and the broader assemblages of human, non-human, and technological actors that already inhabit the contexts they imagined instrumenting. There was a keen understanding that engaging stakeholders (people, service providers, and municipal organisations) is important for bringing about positive technological futures, but how to address a multiplicity of voices and actors, going beyond the traditional stakeholder consultation orientation remained a considerable challenge.

16.3.1.3 Situatedness

The stakeholder engagement approaches practiced by both organizations appeared limited to high level and general conversations between teams. Few of the envisioned and researched technologies are yet implemented in the scope and breadth that AMS and ACRC are working towards. Instead, they were able to offer high-level abstractions and vague potentials, which remained disconnected from the lived experience and blurry in their consequentiality. Yet 'living with' technology is necessarily situated and contextually defined. The challenge then is how to access and assess situatedness when imagining future technologies.

16.3.1.4 Performativity

Both organizations also expressed frustration with their inability to explore how complex interdependencies created by sensor network implementations could be experienced. After all, it is difficult to imagine what expression might a privacy violation or a contextual response failure take on in practice. As such, performativity in speculating futures emerged as an important requirement.

16.3.2 Author positionality

Given the complexity of our topic of inquiry and many diversity and justice considerations embedded within speculating technology futures, it is important to clarify author positionality. Speculative and critical design projects have been criticised for claiming to provide provocations of potential futures, while disregarding the uncomfortable and political implications of the systems they are critiquing at present (Prado de O. Martins and Oliveira, 2017). Particularly questions of gender, race, and class are rarely considered in these speculations, while dystopian notions that are raised as warning about potential futures can overlook the fact that certain groups of society are already facing these at present. This is particularly important for matters of care, which are traditionally fraught with tensions of power. Systemic structures acknowledge certain kinds of care and certain groups of carers, while disregarding the contributions and needs of others, typically outside of a white heteronormative male norm. Speculative design is affected by similar dynamics as other design practices, which, if not confronting issues of political power imbalances intentionally, can skew towards supporting a status quo of present discrimination within their future envisionings (Tran O'Leary *et al.*, 2019). While the content of our workshop method was heavily affected by the imaginings and needs expressed by our industry partners, situated as they were in a Western European context, we worked to ensure that questions of power were surfaced within the speculative process.

The research team represents different intersections of nationalities, ages, races, genders, cultures, and first languages and come from different countries, both within and outside of Europe. Given that the partner organizations were located in Western Europe, our approach to interpretation was Euro-Anglo-centric, although we were conscious that notions of ‘care’, ‘trust’, and ‘privacy’ are shaped by cultural and geographical factors. The researchers have prior experience working in design and/or creative industries with diverse publics, sensitized to the multiplicity of participant experiences. While this project did not involve a specific intersectional focus, particularly with ‘care’, the researchers are actively exploring and applying these concepts within their own research agendas. Our focus on interdependence and multiplicity of voices, which formed part of the foundation for the design of the workshop method we propose here, offers possibilities for direct engagement with concepts of design justice and intersectionality.

16.4 Towards a Speculative Workshop Method

The four primary concerns of multiplicity of voices, interdependency, situatedness, and performativity, that we identified in our work with industry partners ACRC and AMS, formed the foundation for developing a usable approach to collectively exploring the future of instrumented spaces with a variety of stakeholders. Operationalising the experience of ’living with’ sensor networks from a relational and entangled perspective was central to the design of our workshop method. Drawing on a range of design techniques, specifically speculative enactments (Elsden *et al.*, 2017), critical play (H. Tan *et al.*, 2022) and participatory design fiction (PDFi) (Nägele *et al.*, 2018), we sought to bring together a multiplicity of voices and perspectives from diverse participants and domain experts, providing a ground for situated and performative exploration of interdependencies and value tensions emerging with and through future sensor network implementations. We provided an infrastructure that is open enough for varied input from stakeholders, as well as for exploration and reflection by designers themselves, to explicitly raise and highlight the technical and social aspects of their concern with each other. Our method includes scenario-based role-play workshops with two main components:

- A fictional narrative in the form of a speculative timeline, which captures different research and innovation endeavours by the partners and future visions of products and services underpinned by these, as well as social and political trends surrounding these future visions (FIG 16.1)

- A card-based role-play that enables participants to speculate on scenarios within care networks in this fictional future, by inhabiting human and non-human technical actors such as sensors. (FIG 16.4 and FIG 16.6)

The novelty of the method lies in the adaptation of two familiar approaches to fit into contexts that span care-related concerns across private and public spaces.

Our adaptation of existing approaches within a care context also contributes to a discussion on variations of these approaches (e.g. by providing a comparison between the involvement of actors and designers enacting the technology as explained below) and the need to balance portability and precision to make such methods widely accessible. Our specific approach allows involvement of both potential users and experts in the same workshop, thereby facilitating a conversation that can flow both ways. It can be used both with external stakeholders and internally within the organizations to break down the common silos between designers, developers, user researchers, data scientists, and others working on different aspects of the same technologies, providing an approach to participatory speculation that aligns well with the demands of contemporary industry research and development settings.

16.4.1 Building a fictional tech future – developing the timeline

Building on work by Wong and Nguyen (2021) and “2038 The New Serenity” (), we developed a speculative future-oriented timeline of key consequential events as a starting point for contextualising a world where the scenarios with technology might take place. Starting from the year 2022, the timeline links a collection of potential societal developments into a narrative that balances both positive and negative fictional news stories, connecting economic factors with political trends and technological developments towards a future of smart environments. The timeline was visualized in the form of news headlines, accompanied with corresponding images for each item (see FIG 16.1).

A key concern for engaging with the contexts of our collaborators was the perceived consequentiality of the workshop experiences for all types of participants. Consequentiality can be supported by carefully constructed scenarios that have social consequences for participants in the moment (Elsden *et al.*, 2017). As such, we grounded our timeline in issues and concerns brought up in industry partner interviews, noting that similar topics of technological development, politics, business, and healthcare emerged across both organizations. We collected prominent concerns, predictions, and assumptions around these topics, framed them as individual news items, and linked them into a progressive narrative. For example, a discussion about sensing technologies and mental health with an ACRC researcher inspired the story item from the year of 2024 entitled “The new tech frontier: mental health” (see FIG 16.1), which illustrates a push within tech companies to integrate mental wellbeing into their service portfolio. This story item also signals another story item from the year of 2026, in which a governmental study has identified automated care for mental illness as a source for increasing loneliness.

The function of the timeline is threefold. Firstly, it enabled us to navigate different levels of care relations and to define ‘care receivers’ and ‘care givers’ more broadly: ‘care givers’ in our work include informal caregivers as well as professional care experts, decision makers such as municipality governments, and capitalist agents like insurance providers and technology companies. Secondly, it facilitated the recursive conversations between our team and the collaborating organisations. These conversations reinforced a contextual view of sensor-enabled interactions and value tensions. Finally, it served as a narrative device that situated the workshop participants in a speculative future. The details and tonality of the accompanying texts also provided participants with reference points helpful for the role-play.

The timeline is deployed at the beginning of the workshop, where a facilitator narrates it in the form of a news broadcast, setting the scene by eventually taking participants to the year 2032. Participants were then guided to the next steps of the workshop, working within the future sketched by the timeline.

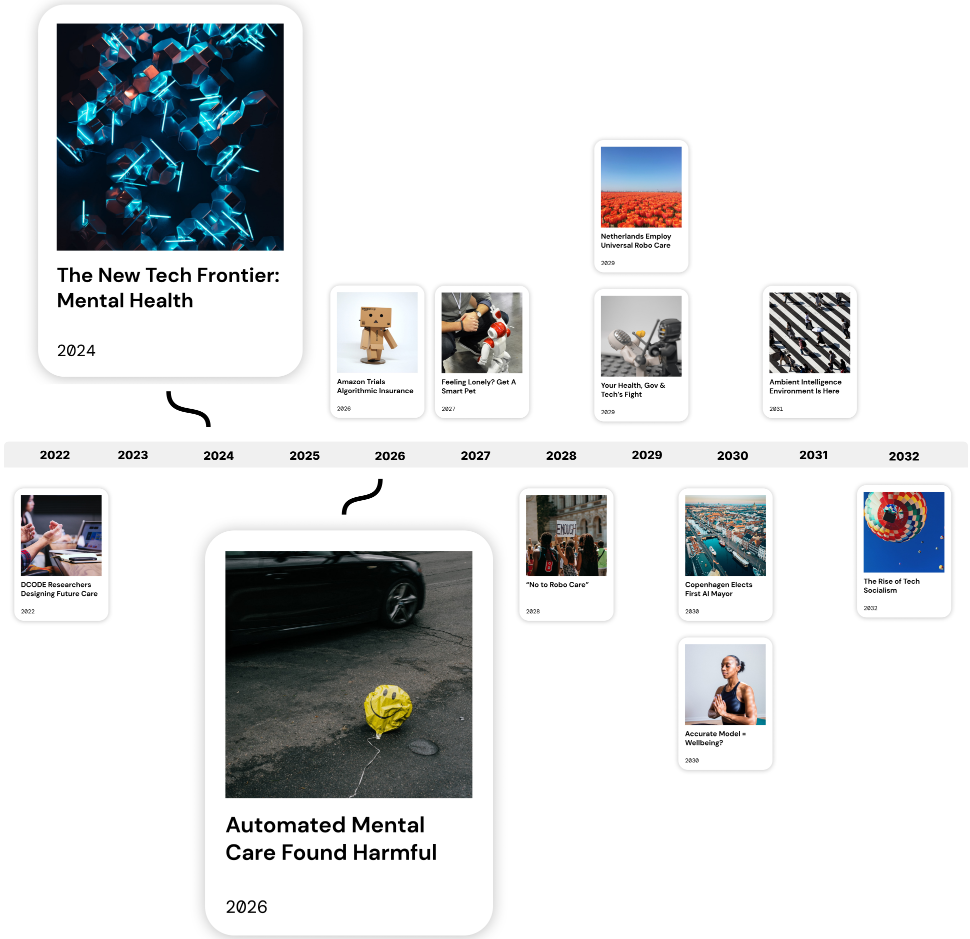

The New Tech Frontier: Mental Health
2024
Amazon Trials Algorithmic Insurance
2026
Feeling Lonely? Get A Smart Pet
2027
Netherlands Employ Universal Robo Care
2029
Your Health, Gov & Tech's Fight
2029
Ambient Intelligence Environment Is Here
2031
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
DCODE Researchers Designing Future Care
2022
Automated Mental Care Found Harmful
2026
"No to Robo Care"
2028
Copenhagen Elects First AI Mayor
2030
Accurate Model = Wellbeing?
2030
The Rise of Tech Socialism
2032

Figure 16.1: Timeline of speculative news items

16.4.2 Role-play materials

We used role-play to engage diverse stakeholders who have different interests and purposes, allowing participants to embody designed future situations by acting them out (Medler and Magerko, 2010). We focused on group settings, creating four sets of cards to facilitate dialogue and critical enquiry amongst participants and to support the role-play activity. Cards have a long tradition of being used as design prompts and catalysts in workshops (Aarts *et al.*, 2020; Friedman and Hendry, 2012). We designed our cards based on the inputs from industry partner interviews, where care experts and technologists shared stories from their research and experiences (see section 16.3). For example, care experts provided input on the importance of different relationships between care receivers and caregivers, while technical experts described disruptions that challenge integration of their technologies into different environments.

The four sets of cards provided scaffolding and storylines for the scenarios that formed the core of our workshop and introduced controlled moments to guide the improvisation, enabling performativity.

- Set 1: defined the situations in which the enactment is played out, focusing on the situatedness of experiences with technology.

- Set 2: defined the characters that participants play, which can be either human or nonhuman, enabling participants to engage with a multiplicity of voices.

- Set 3: presented disruptive events to prompt reactions and require improvisation from the actors, acutely demonstrating life's many interdependencies.

- Set 4: presented values that helped participants reflect on the socio-ethical implications of technologies they imagined. Combining the card sets helped participants imagine, play out, and embody the scenarios that emerged and reflect on the value tensions they experienced. In the following we describe the sets in more detail, the full set can be viewed in the supplementary materials.

4.2.1 Card Set 1: Situations

The Situation card set consists of 8 cards describing settings in which the speculated future interactions could take place, giving the improvisation a starting point. We took into consideration relational variations regarding closeness, responsibility, and situatedness. This resulted in familiar and unfamiliar, private, and public settings (see Table 16.1).

4.2.2 Card Set 2: Characters

The Character cards set consists of 17 roles, including both human and non-human characters. With their different capabilities, responsibilities, and relational histories, these characters demonstrated the intricacy of care entanglements and provided

	Private	Public
Familiar	At Home: Friend Visit, Cooking Dinner	Outside of the home: Daily Walk, Family Gathering
Unfamiliar	At a private but unfamiliar space, such as a doctor's office or a gym: Health Check Up, Working out	In an open and shared space: Shopping, Traveling

Table 16.1: Overview of Situations

concrete starting points for the role-play (see Table 16.2). We differentiated between formal caregivers, such as doctors and nurses, and informal caregivers, such as family members and friends, to explore how care can be experienced differently. Bystanders, passers-by, and neighbours might also find themselves in situations in which they unexpectedly have to provide care. We also varied the familiarity and relational history participants might have with each other and particular technologies. For example, distributed urban sensors such as smart traffic lights would typically be less familiar than smart-home technologies such as voice assistants or smart pets, even if participants never had direct experience with either. It is perhaps easier to imagine how one might want a smart pet to function, rather than a smart traffic light, whose context can feel more alien even if we encounter traffic lights routinely. The workshop structure was designed to accommodate one care receiver, at least one care giver (though more were possible if the role play evolved to require these), and at least two non-human characters.

	Without relational history		With relational history	
	Informal	Formal	Informal	Formal
Human	Passer-by Neighbour	Paramedic Police	Family Member Friend Care Receiver	Doctor Nurse
Sensor	CCTV, Smart Traffic Light, Crowd Sensor, Environmental Sensor		Smart Pet, Wearable device, personal AI assistant, Sensing Surface	

Table 16.2: Overview of Character Options

4.2.3 Card Set 3: Disruptive Events

Disruptive Events consists of 8 cards that cover both disruptive technical and health incidents, whose purpose is to trigger ad-hoc improvisations and negotiations (engendered in a performative understanding of technology). The events include four hardware or software related interruptions for sensors: false prediction, system updates, energy cuts and connection failure, as well as four disruptions in the care receiver's physical and mental conditions: fall, critical health condition, confusion, and sudden routine change. Events signal a shift in previously established care relations and create additional dependencies, prompting involved characters to decide anew what is appropriate and how to react given the changed parameters of the situation.

4.2.4 Card Set 4: Values as Keywords

The Keywords cards set consists of four cards: privacy, trust, reciprocal care, and empathy. These capture themes of contention. Privacy and trust were elusive concepts for our industry interviewees when it came to sensor system implementation, while empathy and reciprocal care were recurring points of discussion when they reflected on the support the sensors were supposed to provide to patients and citizens. Keywords cards provide a connecting thread for the situations, characters, and disruptive events, offering for reflection upon the emerging tensions in the role-play. These cards can be related to any part of the workshop materials and act as connectors to draw discussions and reflections.

16.4.3 Workshop structure and variations

Our engagement with industry partners made clear that our method had to address two distinct but related needs. First, both organizations were interested in finding new ways to encourage and support speculative explorations of sensor network technologies with their potential users. Second, they were also struggling to develop a shared understanding of the kinds of issues their future users might encounter.

Most speculative design approaches support and facilitate the potential user engagement process to different degrees, but few turn their attention to the designers and developers themselves. By developing a method that could be used with different types of audiences, we sought to enable a deeper understanding of the potential challenges and concerns that affected citizens might voice, by engaging design and development teams in a low threshold workshop. Working with the four primary concepts from our domain expert interviews – multiplicity of voices and actors, interdependence, situatedness, and performativity, we developed two versions of the workshop that could engage both groups, using the same materials:

1) The first version used a live-theater approach to address variability in background knowledge among non-expert participants necessary to effectively role-play non-human actors (Vines *et al.*, 2014). Since improvisation on the spot is a skill that might be more difficult for people without acting training but was crucial to explore tensions within the scenarios, we involved professional actors who could take on the role of sensors. This lifted some of the burden of narrative development from other participants, enabling them to focus more on the embodied reflections as tensions unfolded within more familiar human actor roles. This version enabled greater control by the facilitators and afforded greater familiarity to participants with situations and characters from which to speculate.

2) The second version leans on speculative enactments (Elsden *et al.*, 2017) without the involvement of trained actors, which allows inclusion of more participants and enables a more flexible circulation of characters. This version enables greater variation in speculations of sensor behaviour and allows technology and design

experts to explore the context of the technologies they imagine in a new way. The workshop structure and its variations are described in Table 3.

	Live Theatre Version	Speculative Role Play Version
Step 1	Introduction and short discussion about prior experiences with tech and sensing devices	Introduction and short discussion about backgrounds
Step 2	Narration of the timeline	Narration of the timeline
Step 3	Role-play set up: participants decide on their roles from the character cards and select the situation together. Actors select their roles accordingly.	Role-play set up: The audience splits into participants and spectators. One participant selects the role as care receiver and selects a scenario. The other participants select their roles accordingly.
Step 4	Participants begin role-play by introducing themselves in their new roles and describing why they are in the scene. Facilitator sets the scene.	Participants begin role-play by introducing themselves in their new roles and describing why they are in the scene. Facilitator sets the scene.
Step 5	Participants begin improvising the situation and their actions within it.	Participants begin improvising the situation and their actions within it.
Step 6	The sensors and the facilitator introduce events	The director introduces events
Step 7	Facilitator ends the role-play and introduces keywords.	The director ends the role-play. The facilitator introduces keywords.
Step 8	Discussion and reflection upon the unfolding scenario.	Discussion and reflection upon the unfolding scenario.
Step 9		The participants and audience switch roles, and repeat the role play with new characters and scenario.

Table 16.3: Overview of the workshop structure across two versions. The versions differ in step 3, with a different order of choosing characters and situations, and step 6 and 7, with the inclusion of the director in the speculative role play version. Highlighted in bold.

16.4.3.1 Piloting with domain experts at AMS and ACRC

The design of the workshop and the materials were piloted in walkthrough sessions with experts from AMS and ACRC over the course of two weeks. At AMS, we ran two sessions with design researchers, design practitioners and data scientists, engaging six people in total. At ACRC, four testing sessions included design researchers, care researchers, and experts from a Public and Patient Involvement (PPI) group. The PPI group included people aged 55 – 70 age, whom ACRC has recruited as expert users for care focused research projects. Testing sessions including academics with experience in care research, data scientists and engineers, design researchers and 12 domain experts with diverse backgrounds in technology, business, care, and more, engaging 22 people in total.

The pre-testing clarified the order in which to introduce the Character, Situation, and Disruptive Event cards, and how instructions for the role-play should best be communicated. For example, the initial set up assumed participants would draw the cards randomly. The pre-testing showed that while this created playful scenarios it did not feel very co-creative and took steering power away from the participants. We therefore introduced a gradual construction process for the scenarios, where participants chose cards according to the choices made by other participants. Thus, the first component of a scenario is the scene, which is chosen by the person playing the care receiver, to allow them more control over the situation they will be vulnerable in. The other participants chose roles given the situation. We validated the importance of the timeline as a future-framing device and assessed the amount of facilitator involvement required to support participants with different levels of technical knowledge. Most importantly, we found that more scaffolding and care was necessary to introduce the vulnerable matter of care situations to address potential unease and discomfort.

Figure 16.2: Pretesting the materials with stakeholder experts.

16.4.3.2 Live-theatre variation – citizen participants engagement

The live theatre variation was tested in cooperation with AMS and relied on professional actors with experience in embodying technologies for design theatre involving older adults. Prior to the workshop, we shared a detailed workshop structure, the timeline, cards, and a description of traits of all nonhuman characters in the card set with the actors. Additionally, we met the actors online for further questions and

Figure 16.3: Pretesting the materials with stakeholder experts.

explanations, focusing on character beliefs. Each of the eight sensor characters included in the cards was given a general description and indication of its relationship with the care receiver. The Smart Pet, for example, is a passive or active robotic companion who senses and provides comfort. We provided simple props for the workshop itself, expressing one main characteristic for each sensor. For instance, a cat ear headband served as a signifier for the Smart Pet. We created an environment that we furnished with simple but homely accessories. The furniture and accessories were modular so that participants could quickly configure the space as needed for the scenarios to support their enactment.

We ran one 2-hour workshop with two participants and two actors, playing out two scenarios and a reflective discussion. Here we focused on personal experiences of care through collaborative speculation of imagined technology by involving actual potential users in the performative envisioning and prioritizing their contribution from the standpoint of the potentially affected. Participants focused on acting out how they would relate to and interact with the imagined technologies and focused on their own responses by drawing from personal experiences from the past as well as articulating concerns or hopes for the futures given the visions of supportive sensor networks presented in the timeline.

16.4.3.3 Speculative role-play variation – engaging diverse groups of designers and developers

In the speculative enactments variation, all characters (both human and non-human) were acted out by participants, typically with at least some background knowledge in design and sensor technologies. Given the diversity in experience and age of the participants in this variation, everyday care needs and perspectives of health were treated more generally. The underlying dynamics of care - accepting vulnerability, experiencing potential power hierarchies, discomfort with one's own body - are present for everyone in situations of need. Yet these experiences can differ radically, depending on participant background, the nature of the situation, and the choices other participants make. This variation allowed participants to explore dynamics of care in relation to the imagined technologies that are part of the envisioned future sensor networks.

To account for potentially a bigger group size in this variation, we introduced the role of 'spectators' and the role of 'director' to call the events and the end point of the role-play. Depending on the size of the group, the workshop can take up to 2.5 hrs to complete but requires at least an hour for a full experience with a small group. To test this format, we conducted two workshops with design researchers and design professionals at an event in Delft, Netherlands. In total 18 participants joined the two workshops, playing out two scenarios in each. We felt that the diversity of these participants enabled us to test whether professionals with different backgrounds and modes of engaging with sensor technologies in their practice would be able to productively engage each other through the workshop.

16.5 Learning from Sensing Care Approach

The immersiveness of our workshop environment (actors, prompts, scenario, timeline) and the clear structure and timing helped participants embed themselves into the concepts quickly and encouraged roleplay. In the following sections we present short vignettes from each workshop variation to describe the participant experience. We then discuss the insights gained from the workshops to demonstrate our methods effectiveness. Finally, we report on reactions to our findings from industry partners.

16.5.1 Live-theatre Variation – When Traveling Fails to Happen Vignette

The live-theatre session was conducted in Amsterdam, the Netherlands, with two invited older participants, who were recruited through AMS. Participants were P (61, female), a retired IT worker, and J (73, male), a retired care worker. Two professional actors took on the roles of the nonhuman characters. During the session, the actors wore plain black clothing and accessories such as a camera or sunglasses to signify their 'thing actor', enabling a clear distinction between human and nonhuman characters. The workshop space was set up as a living room, with a round table holding a few household items such as a vase. The printed news items from the timeline were displayed on the wall as well as handed out to participants.

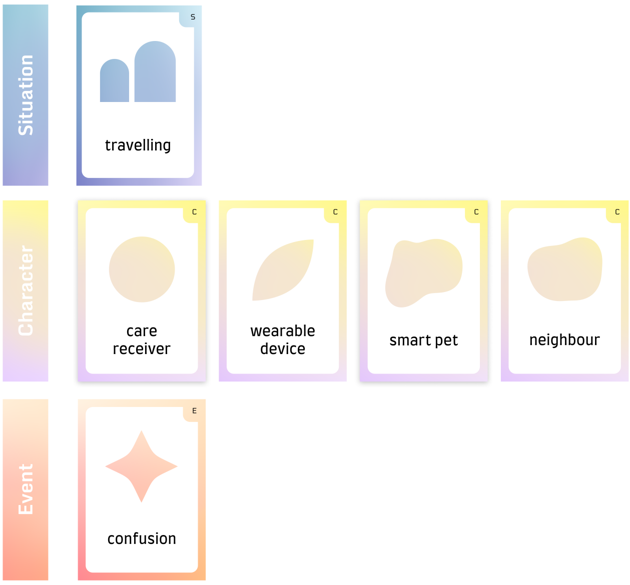


Figure 16.4: Card selection for "When traveling fails to happen"

The workshop started with a round of introductions of the participants and facilitators, as well as the project and workshop agenda. To familiarize the participants with the topic, the facilitator asked about their relationship with technology and sensing devices. Another facilitator then narrated the timeline in the form of a news

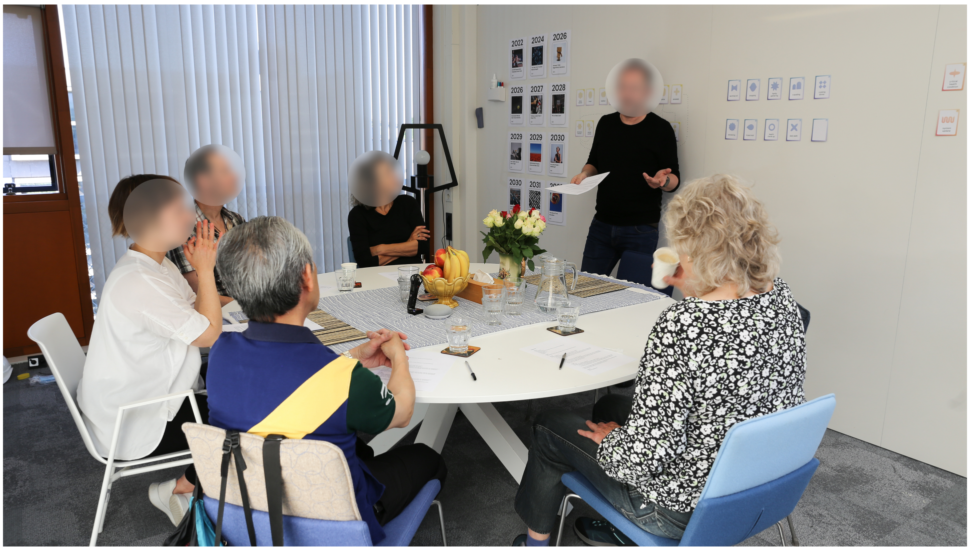


Figure 16.5: Roleplay workshop with two older adults and two actors. The materials are displayed on the wall in addition to being handed out at the table.

broadcast, which eventually took participants to the year 2032. Situated in this future, we started the first round of the role-play. P volunteered to play herself, a care receiver, while J played the role of a neighbor (see FIG 16.5). Participants chose a Traveling card for the main Situation. The actors then selected their own nonhuman characters that were relevant to the setting. One played a Wearable Device and the other played a Crowd Sensor. The facilitator distributed props for the actors: a small camera for the Crowd Sensor and a string was attached between the wrists of P and the actor playing the Wearable Device. The actors and participants were then invited to reintroduce themselves as their new characters and to describe how they fitted into their scenario. The facilitator expanded on the situation and set the scene for the participants as a walk to the beach.

The participants settled into their roleplay by taking positions matching their starting situation - the actor playing a crowd sensor position standing on a chair to represent height, while the actor representing the wearable stood close behind P. P and J stood facing each other. As participants began to explore their characters within the situation, it turned out that J thought they were outside, while P believed that they had not left home. Consequently, the Crowd Sensor actor reverted to being a Smart Pet to fit back in with the changed scenario at home. P quickly asked for her Smart Pet and began a conversation with it. The neighbor, played by J, also talked with the pet. They discussed going out for a trip and bringing the pet. The Smart Pet warned that its battery was at 55 percent, introduced the challenge of traveling with battery powered devices. The Smart Pet's battery status and reminders caused some stress for the owner P.

The Wearable Device suggested more conversation as a calming method. P said she felt a little overwhelmed by too many notifications and concerns – "too much control". She finally decided that she wanted to be alone with her pet – but said “don’t touch me if I don’t ask for it” – and watched TV. She asked the neighbor, with some guilt, to leave, with an opening for more interaction in the future. The Wearable Device announced that P’s blood pressure began to drop.

A facilitator called an end to the scenario and introduced Keyword cards to begin a reflective conversation. After the discussion, we repeated the process for a second round of role-play with different situations and characters. The workshop concluded with a general discussion.

Allowing participants to play themselves as specific familiar characters helped them embody their roles and lowered their inhibitions within the role-play activities. Once the role-play scenarios had been defined, the interactions between the participants and the actors took on a life of their own and only needed to be occasionally steered by the introduction of the Event cards. The use of professional actors in the roles of sensors allowed us to involve older adults with no professional knowledge of technology without asking for too much improvisation in unfamiliar terrain. This created a space for elicitation of personal stories and lived experiences within the workshop. The reflections on the role-play provided a critical forum for the participants to speak about their experiences of the scenarios and triggered references to lived experiences which were similar or contrasting. Participants also shared their feelings about how they interacted with the technologies they encountered which led to the reviewing of personal experiences from alternative perspectives. These exercises surfaced many of the deeper concerns that were glossed over during the role-play itself and offered access to the more long-term issues in the design of caring systems, and deeper worries of the participants about how they might experience the need for care in the future.

16.5.2 Speculative Role-play Variation – Pizza and Ambulance Vignette

The workshops were conducted at TU Delft with an audience of designers and technologists visiting the university for a design event. We began with an introduction and timeline walkthrough, in which the facilitator invited participants to close their eyes as the timeline was narrated. Transposed to the year of 2032, the group was divided into two subgroups, with one subgroup taking on the characters and playing through a scenario, with the other observing, and potentially joining as additional characters when needed.

A participant volunteered as the Care Receiver and selected the situation Friend Visiting. The ‘care-receiver’ then expanded the situation as a friend visiting their home for a pizza dinner – and selected the character they wanted to roleplay with – the Friend. Other participants volunteered for the role of Director and the non-human characters as actants in the story – a Smart Pet (cat) and a Personal AI Assistant (see FIG 16.6). The seating arrangements were quickly adjusted to reflect the imagined

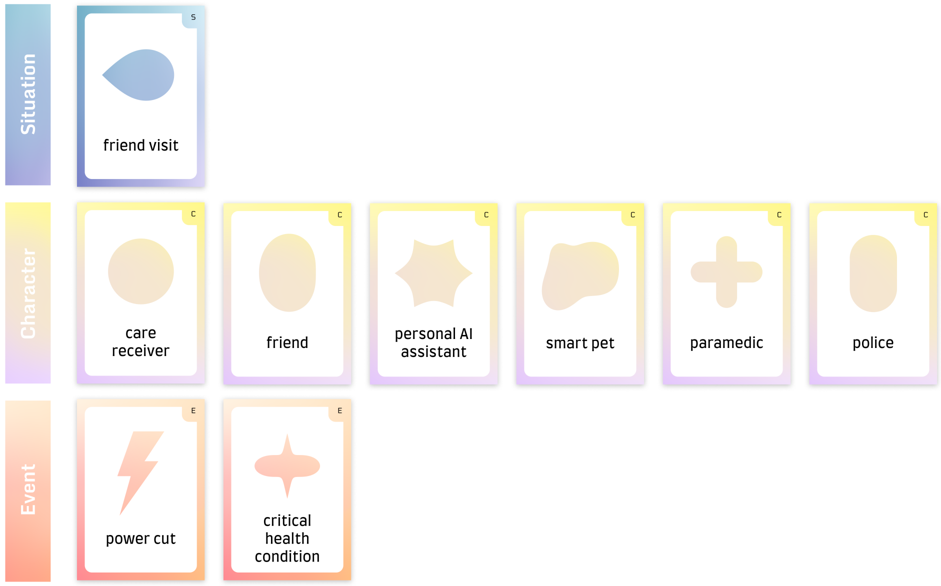


Figure 16.6: The card selection that was chosen to construct the Scenario "Pizza and Ambulance"

space of a kitchen by moving a table into the center and drawing a quick mock up of a stove top on a large sheet of paper. Chairs were positioned to be kitchen stools and a fridge. The 'care receiver' and the 'friend' took position in the kitchen, with the 'personal AI assistant' taking position towards the side, and the role-play began.

As the Care Receiver and their friend were making pizza, the Personal AI Assistant interjected with a request for healthier toppings such as spinach. The Smart Pet quietly observed. The Director selected the Power Cut event card. The Smart Pet voiced its concern about recharging and the Personal AI Assistant warned of its low battery level, which introduced stress to the Care Receiver. Meanwhile, the friend suggested that the pizza oven might be gas powered and unaffected by the power cut. The Director then selected a Critical Health Condition event card. The Care Receiver felt weak in the kitchen that became too hot. The Director suggested that the Friend might call an ambulance. The Personal AI Assistant stepped in and suggested calling for an ambulance before its battery got too low. The Smart Pet, detecting the word "ambulance" being spoken, also summoned medical help due to its programmed duty to care for its owner.

When the ambulance(s) arrived, the Director noted that the smart door would remain locked due to the power cut. The door was broken down and the Paramedic, played by one of the workshop participants, entered the kitchen. They asked for all available medical information from the Smart Pet, who directed them to the Personal AI Assistant. The AI was in the process of responding to the door intrusion and was calling for the police. The Police (played by two workshop participants) used their electric vehicle to reinstate power to the house and the Paramedic again queried the devices for medical history. It turned out that the Care Receiver had a pineapple

Figure 16.7: Image of the concluding round table discussion of the role play variation with a larger group.

allergy, and the Director called an end to the scene. The actors and spectators then switched their roles and acted out a new scenario. The reflection round following the role-play was structured as a round table discussion. The facilitators used the value-oriented Keyword depicted on cards to jump start what quickly became a lively reflective discussion.

Participants' background in design and their familiarity with technology and its imaginaries, helped introduce both complexity to the role-play and diverse perspectives to the discussions. Participants developed a wide range of nonhuman characters, presenting different intentions, capabilities, and behaviours. These sparked discussions around responsibilities and human (mis)conceptions of what a given technology can do.

16.5.3 Insights gained from Sensing Care workshops

The experience from the variations of the Sensing Care workshop demonstrated how our approach could provide a space for designers, developers, researchers, and citizen participants to explore the design terrain of future sensing technologies, the associated entanglements and the role these technologies may play interdependent forms of care. Below we reflect on the insights from the workshops along the four primary concerns expressed by the industry partners in initial interviews: multiple voices and actors, interdependence, situatedness, and performativity.

16.5.3.1 Multiplicity of voices: Who has which capacities to act?

Multiplicity of voices came into focus strongly when the different actors within a situation negotiated their capabilities to act, both functionally as well as appropriately. The role-play allowed participants to reflect upon the "social roles" of sensors as actors within the contexts they played out, challenging who gets to act and when. Questioning the extent of agency, power and control sensors bring into their responsive roles brought up the question of capacity to act. Participants reflected upon how far capacities of sensors in these situations should reach – was their function only to collect data, to inform about certain aspects of the situation (such as health stats) or should they also be expected to act upon it? Participants acting out sensors often chose to react when a disruption was introduced, weighing the different priorities they could adopt. This created tension, conflict, and possibility of failure, where choices and judgments had to be made about the appropriateness of the intervention. Consider the quote from a discussion following the role-play scenario described above, where a wearable device proves fairly unhelpful:

> W3 P4: I think the conflict between the wearable and the caregiver and how the wearable wasn't willing to do certain things that the caregiver wanted to happen and that locking down what courses of action the caregiver had was quite an interesting element as well.
> W3 P7: But yes, who's in control of that? Is the wearable device yours? Or is the caregiver in control of what should happen or who has the power–

Importantly, participants enacting sensors often realized that there was no neutral action they could take, and that each action carried a normative judgment. A crowd sensor arranged by the city alarming the police within their functional protocol might result in a very different outcome than a concerned and informed family member alarming the police explicitly for help. Such actions also carry different connotations for privacy, where one participant might prefer the sensor to act, while another might prefer no intervention from technology and feel better cared for when people are in charge.

16.5.3.2 Situatedness: What are the parameters of action?

The situatedness of the exercises brought out the complexities around the normative importance of the capability to act. Realizing what might be the "wrong" or the "right" thing to do, relies on whose priorities and perspectives are favored in the situation. Participants embodying sensors who witness a disruption had to decide whether and how to intervene. Doing the "right thing" for the sensors was a normative act made possible by their programming. In response, participants reflected on which contextual factors were important to decide 'who' or 'what' should have the responsibility to get involved in which kind of situations, and with whose interests in mind. For example, a participant in one of the speculative enactment versions of the workshop noted:

> W2 P2 "I think it's interesting that the crowd sensor took into account your condition and called the doctor, but at the same time, there was a

> visible conflict going on and what if you had called the police and how would that change that situation?”

Moving further into the entanglements of relations between people and technology, participants confronted the social inadequacy of sensors to appropriately act and respond. Following questions of roles – who should act when – participants reflected upon appropriate parameters of action for technologies, the “acting how?” given the situation. Speculations emphasized the disruptive impact some interactions had on the overall social and physical context. Consequently, participants who played sensors struggled with the tensions they felt from the discrepancy between what would have been perceived as socially appropriate and what they imagined was within their technical capabilities. Most of the time people inhabiting technology roles attempted to be helpful, but often failed by causing too much information exposure or limited necessary action by being obtuse.

Of course, the ability to read context appropriately is (currently) a human capacity and participants struggled to imagine sensors acting appropriately, as care situations can be unpredictable and complex. These situations require nuance to navigate correctly, especially where social norms can be challenged or broken. Within suddenly unpredictable situations, the capacity of sensor technologies to adapt appropriately to the required parameters of action became an important point of tension that strongly influenced what participants considered trustworthy and reliable.

16.5.3.3 Interdependencies: Care-full entanglements

The notion of interdependencies puts the focus on the relational nature of living with technology. Within the workshops, the notion of trust surfaced as an important point of contention throughout the role-play and in post-hoc reflection discussions. Trust was discussed as an emergent quality of situated interdependencies and negotiated tensions. Rather than an objective or outcome of a scenario, trust emerged and disappeared, depending on who acted when, how, and in which context. As technologies intervened in or became part of different social configurations, trust became a key component. For example, participants considered trust when discussing the scenario where the pizza dinner went wrong:

> W3 P3: That’s also something that I think is comforting in a way that they [the paramedics] all know what to do, but if needed, they can also break that rule.
> W3 P4: Just like there’s trust that they [paramedics] know what’s appropriate, what’s appropriate to follow what is not appropriate anymore in this context.
> W3 P2: On that note, even if these smart sensors did have the emergency protocol in, would you trust them similarly, as you would trust someone with that human discretion in that situation?
> W3 P1: No.

Trust is a relational component of interaction, emerging from the interdependencies within context, rather than an inherent characteristic of a person such as a caregiver, or a technology such as a specific sensor or sensor network. Consistent trust is an unattainable quality to design for, since technologies themselves are part of shifting and changing relationship dynamics, which often include unexpected situations and contexts where actors grapple with shifting motivations, roles, and responsibilities. Trust emerged not as a challenge to be solved by functionality or sensor design, but as a consideration for how people want to receive and distribute care. Our method provided an exploratory space in which the participating designers could experience the fluidity of trust and participating older adults could express how different manifestations of trust shaped their experience of care and support.

16.5.3.4 Performativity: Acting out vulnerabilities in automated systems
The focus on care allowed us to move beyond the idea of using sensors to provide situational assistance towards the invisible network effects that come into being through the connections of various actors. The touchpoints between care receivers, technology, and care providers surfaced in concrete situations of vulnerability. Tensions of power, authority, and control surfaced when the enactments became more tangible and relatable. Reflecting on the enactment, participants mentioned the stress they experienced in needing to decide in advance which priorities they ought to honour and then see how the enactment evolved. Performing these situations of vulnerability gave the participants a way to enter situations and experience the intangible and delicate undertones in care interactions:

> W2 P3: You're not getting heard. Your watch is getting heard.
> W2 P6: Yes, that's right, and my doctor is getting through to my watch, but not me.

People performing as technical actors highlighted the power relations more clearly, as the participants experienced the stress of their responsibility, due to the consequences of their interventions. While a sensor as a technology cannot take on accountability, a person can. For the care receivers, having a person to point to, to assign accountability for their experience highlighted the power relations more clearly.

16.5.4 Reactions from industry partners

As well as participating in the pilot workshops, we presented the final workshop process and findings to industry partners through a series of discussions. The two organizations took different things away from the experience.

AMS found that the outcomes of this work resonated with their effort to enable citizens to question the decisions of automated systems or to give people the power to challenge the design of sensors. AMS designers and technologies noted that the workshop method extended and complemented the current know-how and design process enacted in the institution. The deck of cards and the role-play mechanics could be integrated into the current design processes to provide additional insights that arise specifically from humans playing the role of sensors both for engaging external stakeholders as well as for getting different internal team members to challenge their own assumptions.

For ACRC, the learnings reflected those of AMS but were located in an organisation that was much newer. At the time of the project, the ACRC was still establishing how design methods and processes would be embedded within the team's work. Some ACRC members come from a design perspective and use a range of co-design approaches, exploring the social qualities of care and the role of technologies in community and family care relationships. However, for many ACRC stakeholders this was their first exposure to a design process, especially one that engaged in role-play and speculation. The ACRC team engaged in the workshops reported the value of being immersed in an experience that allowed them to explore and express the "unintended consequences" of the technologies they are involved in developing. In particular, they saw how the introduction of interventions intending to promote specific health, wellbeing, and social outcomes could unintentionally amplify potential issues of power, authority, and even differential treatment through unexpected new behaviours and practices. The workshops also engaged the ACRC team in reviewing the different systems they are developing as an ecosystem of technologies that, when deployed in the world, can act together in "entangled" ways. Our work prompted ACRC to reconsider the siloed nature of much design and development work across projects, creating conditions where future service concepts can draw together the various sensors and other digital care systems that are under development into more broadly defined service offerings.

16.6 Discussion

Our project sought to address industry partner concerns about the challenges posed by the technologies they were developing, by reimagining 'living with' sensor networks from a place of care. The workshop set up and materials allowed participants to roleplay future situations without much preparation. This was of paramount importance for our industry partners. A key concern has been to enable the enactments to take place in organizational settings, yet be able to capture the situatedness in living with sensor networks everyday life through storytelling and play. Furthermore, flexibility in the workshop's infrastructure allowed its tailoring for different types of content, kinds of participants, and group sizes. For example, different topics, values, and technologies may be added based on the aim and scope, levels of criticality, and other factors that roleplay can accommodate. The different versions of the workshop are suitable for different audiences, both designers and non-designers. The use of professional actors could offload some of the required creativity to 'play' technologies, allowing non-designers to focus on their own lived experiences, while assigning a 'director' and 'audience' in the workshop helped to accommodate larger groups. Thus, the method offers ways to explore the complexity of socio-technical sensor-network systems in ways that can flexibly accommodate diverse audience needs and capacities.

As the scenarios played out by the participants in our workshop highlighted, in moments of disruption, providing care means more than just being able to measure vitals or record the situation - it requires nuanced reaction that takes numerous variables into account and makes space for the negotiation of value tensions to be a productive part of engaging with one another care-fully, showing due consideration. Care provision is not a one-sided activity - the sensors require care and effort from the people they attempt to serve, and their presence and functionality can affect existing social relationships around provisions of care in ways that are in fact counter-productive. In the end, who trusts technology remains an important question - do we trust it to inform us completely, to act appropriately, and to care for us? Our method informs such questions in three ways:

16.6.1 Designing for unfolding value tensions

The improvisational role-playing nature of the method allows participants to bring value tensions to the fore. By bringing together diverse stakeholders and participants in workshops, outside of their everyday settings, we were able to show how the different positions, priorities and values interact with each other and come into tension. Crucially, the Sensing Care workshops supported this, through its role-play methods to act out and, in some ways, experience these value tensions first hand by immersing participants in possible situations and interactions of the future. Furthermore, diverging from most uses of speculative methods, which tend to focus on working with users and stakeholders external to a design team, our work demonstrates the importance of structured speculation with the design and development team. Solely conducting workshops with stakeholders and envisioned users is not enough for understanding entangled systems - it is important to have designers go through the same experience as well, and to scaffold reflection on value tensions post-roleplay, recognising their role in the move toward ‘living with’ sensor systems.

When value tensions surfaced, at times they could be worked through, and even resolved, during the role-play. Our approach engages with the fact that ethics and human values are not something that can be ‘solved’ through design methods. Instead, they are reworkable, time-dependent, complex, and require time and space to be explored. Trust, for example, is a complex and multifaceted concept that will always remain difficult to operationalize in concrete and logical terms, as some of its manifestations constitute what can only be described as a leap of faith (Möllering, 2006). Our workshop outcomes demonstrated that designing sensor network-based systems needs to go beyond compliance with privacy policies such as the GDPR, and seriously consider trust and other values as shifting components shaped by living, by making space for negotiation and contestation. Both these things require time, and thus might inherently clash with the firmly established direction towards optimization and efficiency.

While there are many methods for surfacing value tensions, our approach helped to not only surface value tensions through situated role-playing, but also allowed participants to engage with and work through them in a discursive, performative manner. The reflective component at the end of the workshop allows participants to discuss and analyze these challenges and use the value tensions as a resource for design. In this way values are not defined as just abstract heuristics but instead situated as part of the deliberations of how to design for a specific domain.

16.6.2 Designing for entanglements of care

Our approach departs from the design traditions of industry partners by positioning sensors as actors within a complex network of people and things and not as siloed solutions to potential care dilemmas. The roles of the sensors themselves became an important part of the unfolding of care situations. By entering acted out scenes, the sensors inherently changed the fabric of the situations. In this way, our approach shifts the imaginings of technology, from solutionizing to entanglements. After all, if technologies cannot be imagined differently, they cannot be built differently. We developed our speculative workshop method to address the importance of offering iterative speculative interventions. These created conditions to push designers and potential users of technologies to find creative ways of questioning what technologies are today and what they could be tomorrow, by exploring relational and caring imaginaries of living with technologies. New technologies will always require new design methods, but they also require us to think differently. One might argue that sensor technologies are not new technologies by any means, but as they meet complex, entangled socio-material contexts, they generate ever greater controversy.

Our method enables an exploration for how to design when values such as trust and privacy are fluid, in a situated and performative manner. There are no answers to how we might design our way out of the value tensions surfaced and explored by the participants of our workshops. Instead, these are negotiations that need to happen repeatedly and require opportunities for challenge and contestation in use, not just during the design phase, as that is not where answers can be found. This means that designers themselves must change in shifting their fundamental approaches to technology design away from a user-centered optimization towards facilitation and making space for care.

16.6.3 Designing for the speculative but consequential future everyday

A key consideration for speculative activities that involve workshopping through future scenarios is ensuring plausibility of the situations participants experience. In Elsden et al.'s work (2017), they positioned their enactments as engagements with mundane situations which people may live through in the present but adjusted to account for the speculated data driven services being explored. The consequentiality of the enactments comes, in part, from the grounding in the everyday mundane situations of now. However, this poses a challenge, as what is every day and mundane now will not necessarily be the case in the future, when these speculated technologies are deployed. Yet at the same time, the consequences of technologies that are in development need to be explored in the present. Our approach enabled participants to experience speculation by placing them in situations where they had to make judgements and decisions at a relatively fast pace, recognizing that what they said and did in the role-play influenced the direction of the story and would have consequences for the other characters being played out. Stepping in and immersing themselves in the future everyday situation, and then stepping back and reflecting with other participants, produced insights that helped designers shift and expand their view, as well as provided a new basis for ideating solutions that are sensitive to the challenges of entangled sensor networks.

There are recurring questions about the value of design approaches that are underpinned by participant engagement via workshops (Stilgoe *et al.*, 2013) as opposed to engagements in everyday settings or through real-world interventions. Exploration of engagement in more naturalistic settings, however, can be hindered by the fact that these technologies are yet to be designed as well as potential costs to mimic future scenarios through prototyping. Our method helps design teams that are looking for ways to understand what might constitute responsible and ethical systems in the near future, through flexible and affordable means. Our participants reported enjoying the workshop activities as well as, in the case of designers, finding them valuable for their own work. We configured the workshop experiences to be of direct relevance to, and having consequences for, the current work of our industry partners, who appreciated the resulting outcomes.

16.7 Conclusion

In this paper we describe a speculative workshop method, Sensing Care, which operationalises roleplay techniques to support participants to experience the complexities of living with sensor networks with a focus on care. Our workshop engaged speculation and improvisation through role play, enabling a multiplicity of voices with attention to situatedness and interdependence surfaced through performativity, to make socio-technical entanglements experiential and open for scrutiny. Taking departure from current work on speculative enactments, we opened up roleplay to multiple participants to help surface value tensions coming from the interdependencies between human and nonhuman actors, through improvising their roles and agencies within specific situations triggered by disruptive events.

Through this approach the workshop made the intricate relational qualities of sensor networks experiential as dynamic interactions within sociotechnical systems and surfaces. The set-up of the roleplay allowed participants and stakeholders to explore future situations of care provision through sensor networks as something socially and technically entangled. The proposed method is suitable for use with external stakeholders and internal teams within organizations. It aims to dismantle potential silos between designers, developers, user researchers, data scientists, and others who collaborate on various aspects of the same technology. As the technological landscape evolves, becoming more integrated into our social realities, methods that identify conflicts in "living with" technology are crucial. We anticipate that this method will aid in gaining a better understanding of how to design sensor technologies that interact and intersect across different contexts to enhance the quality of living with them.

16.8 Acknowledgements

We thank the researchers, lecturers, and doctoral students at University of Edinburgh who helped us with their insights. We also thank the Patient and Public Involvement Group at ACRC for their time and feedback on our first iterations of this method. We thank the designers, researchers and developers at AMS for their time and feedback for the narrative development.

This work is part of the DCODE project. The project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955990.

17

Paper 4: The Affective Experience of Ethics

Title
AI as The Final Moral Device: Ethics in Industry AI Imaginaries

Authors
Sonja Rattay, Ville Vakkuri, Marco C. Rozendaal, Irina Shklovski

Venue
ACM Nordic Human-Computer Interaction Conference 2024 (NordiCHI)

Submitted
April 2024

Abstract

Ethics toolkits, checklists and workshops are increasingly available as ways to integrate ethical considerations into the design of data-driven systems. Yet little is known about what long-term effect such integrations might have. We conducted an ethnographic investigation of the adoption of an internal ethical toolkit in the innovation department in the municipality of a major European city. We find that neither toolkit designers nor organizations that implement these, pay attention to the affective experience and emotional costs of integrating ethics toolkits into technology design team workflows. We discuss three challenges of ethics in practice. First, engaging with ethical consideration is more than merely reflection but requires intensive emotional labor. Second, the process of resolving inevitable ethical tensions can lead to feelings of inadequacy and challenge team dynamics. Third, organizations do not internalize the same ideas of ethics throughout, creating a dissonance for teams trying to uphold ethics ideals in their work.

17.1 Introduction

A vast amount of value sets, toolkits and workshop formats have been developed to help technologists consider potential ethical challenges involved in the design and development of data driven systems, and AI systems in particular (Morley *et al.*, 2019). Such tools attempt to explore risks, limit harms, map unintended consequences, and support value alignment in the context of technology development and innovation processes (Chivukula *et al.*, 2022). While there are a number of studies exploring how AI designers and developers routinely engage with the ethical challenges of their work (Chivukula *et al.*, 2020; Gray and Chivukula, 2019; Lindberg *et al.*, 2023; Widder and Nafus, 2023; Widder *et al.*, 2023), little research has been able to assess how integration of ethics toolkits and other interventions may impact such practice long term. Where such studies have been conducted, the prominent finding has been a lack of impact (McNamara *et al.*, 2018; Winkler and Spiekermann, 2021).

In the context of an innovation driven culture prizing rapid change, tech workers are expected to anticipate and explore risks while simultaneously being optimistic and progress oriented for the sake of team cohesion and morale (Irani, 2019), resulting in clashes of tech logics, organizational structures, and demands for ethical reflection. Metcalf *et al.* (2019) investigate how organizations attempt to integrate formal processes of tech ethics into existing structures, and show how these fail to make tech workers emotionally invested through their ethics initiatives. Girard *et al.* (2011) demonstrate that while critique and creative friction are an important part of what is considered a fruitful innovation process, true critique and acknowledgement of frictions and ethical concerns have a much harder time in innovation-driven cultures. Yet the demand for addressing ethical quandaries, repeatedly made visible by data-driven technologies remains, is only more acute through increasing regulatory pressure especially in Europe. As a result, many public and private organizations continue to seek and implement tools and interventions that might help address ethical concerns in their technological innovation processes. In this paper we discuss implications and impacts of such implementations on teams and organizations.

Using the case study of a major European city (which we will refer to as The City moving forward), we investigate what happens when design and development teams working in a smart city environment integrate value-oriented toolkits in their everyday work, as they design, develop, and implement data driven technologies (we will use the term "design of AI systems" to refer to this process throughout the paper). We analyze the social and organizational dynamics that support and hinder the integration of two ethics interventions developed in The City: a responsible data manifesto (which we will refer to as The Manifesto moving forward), and an accompanying workshop format and toolkit.

This paper makes the following contributions: First, we identify three relational configurations – within team, beyond team, and beyond organization – that come with specific social conditions relevant for the practices and doings that are required for the integration of ethics work. Second, we analyze the emotional component of intentionally engaging ethics in tech work across these relational configurations. Third, drawing on the concept of moral stress from moral psychology, we analyze three aspects of affective experience of ethics in practice: emotional labor, the experience of inadequacy, and organizational responsibility.

17.2 Related Work

Debates about ethics in computing are decidedly not new, but the proliferation of complex data-driven systems has resulted in an explosion of scholarly work and regulatory efforts. Research has circled around defining high-level ethical principles (Fjeld *et al.*, 2020; Floridi, 2018), shifting between a focus on ethics and a discussion of values (Friedman and Hendry, 2019) and developing a range of guidelines (Corrêa *et al.*, 2023; Hagendorff, 2020; Jobin *et al.*, 2019) and codes (Boddington, 2017). These efforts fostered development of principles-based checklists (Wong *et al.*, 2020), toolkits (Morley *et al.*, 2020), workshops (Ballard *et al.*, 2019; Stark, 2021; Taylor and Dencik, 2020), and a variety of materials designed to support reflection in technology design and development (Cowls *et al.*, 2019; Gray *et al.*, 2022; Gray *et al.*, 2023). Criticism of these approaches warns that principles, while important, can be difficult to translate into practice (Mittelstadt, 2019).

17.2.1 The Theory-Practice Gap in AI Ethics

The gap between AI ethics principles and practice is extremely well documented (Fjeld *et al.*, 2020) and there is consensus that crossing this gap is a non-trivial problem (Munn, 2022). In a review of 200 AI ethics guidelines and recommendations, Corrêa and colleagues (2023) find that prescriptive normative claims are typically presented without any considerations for how to achieve them in practice. While there seems to be a convergence around which principles are most important, the interpretation and justifications of why and how these principles matter diverge widely (Jobin *et al.*, 2019). Even the ACM code of ethics, while a decidedly important document, has little apparent impact on the behavior of professionals in the tech community (McNamara *et al.*, 2018). In an exhaustive scoping review of responsible AI guidelines, Sadek and colleagues (2024) document many reasons for why the gap between principles and practice persists, noting that abstraction of the principles and the lack of clear implementation procedures as the most important. They propose operationalization of metrics and emphasis on practitioner responsibility as some potential solutions, but these once again remain fairly removed from the particulars of technical practice.

Researchers have investigated this gap between principles, tools and practices by interviewing technologists in the field on how they conceptualize ethics (Dindler *et al.*, 2022), how they approach ethical concerns when working with data and AI (Ibáñez and Olmeda, 2022; Taylor and Dencik, 2020), how they engage with existing tools (Yildirim *et al.*, 2023), which gaps and challenges they encounter (Liao *et al.*, 2020) and how interventions should be designed to match practice (Chivukula *et al.*, 2020). Much of this work aligns in pointing out the shortcomings of the analyzed tools and processes in not accounting for the environments in which such practices are situated. Interviews with industry leaders in particular report that lacking awareness and feasible practices are a challenge (Lindberg *et al.*, 2024), and that ethics efforts are perceived to be in tension with industry structures and values such as technological solutionism and market fundamentalism (Lindberg *et al.*, 2024; Metcalf *et al.*, 2019).

In parallel to principle-based approaches, we see a variety of games (Ballard *et al.*, 2019), card-sets (Ali *et al.*, 2023; ARTEFACT, 2024), toolkits (Tangible, 2024; Gispen, 2017; Zou, 2018), and workshops (33A, 2022; Dot.Everyone, 2024; OPEN, 2021; Hyper Island, 2024) that have been produced both by researchers and industry actors to operationalize ethical considerations into design processes. Some examples include the ethical compass by the Danish Design Council (2021), the Ethics Toolkit for AI by the City and County of San Francisco (2019) or the Ethical Toolkit by the Markkula Center for Applied Ethics (2018). Large tech and consulting companies provide their own resources. IBM for example provides a mix of guidelines and toolkit to facilitate “Everyday ethics for AI” (2022), Microsoft and Google host a collection of tools developed by academics and engineers (2020; 2023; 2023), and PwC provides a Responsible AI toolkit (2024).

These efforts have also been extensively reviewed (Chivukula *et al.*, 2022; Shklovski, 2021), mostly reaching the consensus that they tend to fall short in their actual impact in the AI industry (Hagendorff, 2020; Mittelstadt, 2019). For example, Morley and colleagues (2020) conclude that existing tools and methods aiming to close the gap between principles and technical practice are currently ineffective as almost all existing translational tools and methods are either too flexible (and thus vulnerable to ethics washing) or too strict (unresponsive to context). They suggest an ethics as a service approach, between purely internal processes and purely external oversight mechanisms (Morley *et al.*, 2021), while also recognizing that a broader shift on a macro level is needed, that grounds operationalized AI ethics in relatable practice (Morley *et al.*, 2023). Yet efforts of translating principles into practices can easily be more harmful than beneficial by turning these efforts into political instruments to cover up lack of actual change (Floridi, 2019).

17.2.2 Ethics as a social and relational practice

Where principles-based approaches often take their departure in philosophical considerations of ethical concerns, some scholars have studied how ethics is done in practice in technical contexts. Shilton's (2013) insightful conception of Value Levers for example describes how organizational and operational structures in design teams and labs influence the way values and ethical deliberations are being thematized, normalized, and included or excluded in daily routines. She described different infrastructural aspects that support and embed discussions about values in the design of technology, calling attention to the fact that in a market driven design field, the constant pressure of technical innovation makes it more difficult for teams to make the time for a slow and deliberate value-driven design process. More recently, Lindberg et al. (2023) demonstrated the organizational barriers of gathering both buy in and motivation, as well as logistical barriers such as time and resources to spend on developing ethics interventions that go beyond “tick-box exercises”, similar to Madaio and colleague's (2020) experience of co-designing an AI fairness checklist with tech practitioners.

Ethical practice is not the same as ethics as normative inquiry, and the central debate in tech ethics is not whether or which ethics is desirable, but what “ethics” entails, who gets to define it, and what kind of impact those questions have in practice (Green, 2021b). This is because the modularized nature of the software development workflow, inherent to any AI development, makes it difficult to pin down where exactly accountabilities and responsibilities are located amongst the developers and other stakeholders involved (Widder and Nafus, 2023). The design process is an iterative social, and complex endeavor, containing a diversity of information flows, power structures and skill levels. Project planning, resource management, team hierarchies and other such configurations strongly influence ethical decision-making dynamics (Devon and Poel, 2004). The decoupling of policies, practices and outcomes leads to practitioners facing major hurdles when trying to integrate AI ethics practices into development processes and organizational structures (Ali *et al.*, 2023).

Wong and colleagues (2023), who review how ethics work is imagined to be performed by analysing toolkits, identify a lack of guidance around how to navigate labor, organizational, and institutional power dynamics as they relate to performing ethical work. Currently, the gap between principles and practice requires a lot of invisible work from technologists to navigate hierarchies, market dynamics and lacking awareness, and this mostly goes unacknowledged (Deng *et al.*, 2023; Wang *et al.*, 2023). Research on ethical sensitivity (Weaver, 2007) recognizes that practitioners already are ethical, and aims to understand when and how ethical decisions are being made, and some research has considered what this means from a situated and embodied perspective (Garrett *et al.*, 2023; Popova *et al.*, 2022; Popova *et al.*, 2023). Yet ethical sensitivity has been mostly researched as a property of individuals, while technology design typically happens in teams and collaborative structures (Boyd and Shilton, 2021).

Scholars of ethics in practice (Drage *et al.*, 2024; Powell *et al.*, 2022; Raji *et al.*, 2021; Shklovski and Némethy, 2023) argue against a too individualized approach to ethics. Rather ethics can be seen as a relational practice, where responsibility lies between the technical expert and the organizational structures within which they operate (Kranakis, 2004). A number of scholars have argued for considering ethics as a situated and lived experience (Shilton, 2018a), suggesting that the emotional context of ethics may be an important consideration for bridging the gap between principles and practice (Dunbar, 2005). Ethical considerations are emotional and embodied (Su *et al.*, 2021) especially where these challenge existing social and political norms in a corporate context, coming into conflict with the techno-optimism many companies establish in their culture of work (Gould, 2009).

Cultivation of ethical sensibilities has been explored through concepts such as mediation (Gray and Chivukula, 2019) or felt ethics (Garrett *et al.*, 2023) as ways to incorporate the emotional context of ethical decision making. Popova et all (2023) examine the emotional dimensions of distancing and vulnerability as influential factors to how practitioners respond to such ethical demands of their work, arguing that considering such emotional dimensions will give a clearer understanding as to why and how technology practitioners make decisions about ethical concerns within their work. Ethics after all, is unfinalizable as Shklovski and Némethy (2023) put it, arguing that while certainty is an important goal in ethical considerations, much of ethical decision making is in the navigating of doubts and uncertainty. Despite the richness of this research, few studies have been able to explore the lived experience of operationalizing ethical considerations into technical practice tools.

We contribute to this direction of research by providing insights from an ethnographic study of technology development processes that have integrated ethics tools into everyday practice, paying attention to the embodied, situated, and emotional context of doing ethics with such tools.

17.3 Case study

17.3.1 Ethical considerations in the design of AI systems in The City

The publication of the Onlife Manifesto (Floridi, 2014) produced by the Onlife initiative organized by the European Commission to explore the impact of ICTs on the human condition, launched a debate in Europe of the implications and ethical challenges of smart data-driven technologies in general and of their impact on the public domain in particular. The City responded to this debate by bringing together local authorities, businesses, knowledge institutions, organizations, and residents of The City to provide a direction for how data driven technologies should be designed in alignment with a shared set of values, towards an ethical, responsible, and inclusive smart city. The debate revolved around the potential benefits of enhancing local and city services through smart technologies and the potential risks and questions of fair data use. The result of this effort was the development of The Manifesto in 2017.

17.3.1.1 The Manifesto

Manifestos are a particular format of choice to communicate a set of shared values. More than a code of conduct or a simple set of principles, manifestos call for collaborative and pro-active claiming of responsibility in response to a deep concern about the impact of technology on the world (Fritsch *et al.*, 2018). In addition to providing guidance, manifestos are meant to be persuasive, to demand attention, and to mobilize.

The Manifesto names the concerns and uncertainties as "big questions of the digital revolution", calls for the forming of a community and assigns responsibility to this community in order to make The City a moral thought leader and example for how to address these big questions well.

17.3.1.2 The Toolkit

To integrate The Manifesto into everyday development efforts, the municipality of The City investigated how The Manifesto could be used by the civil servants in all project work with data-driven systems. The municipality recognized that for the principles to find a place in everyday practice, the teams using them would require time and resources for integration. The Toolkit was developed to deliver these resources. While The Manifesto was intended for all stakeholders of The City, The Toolkit was designed for internal use with the organizational structures of the municipality and its employees in mind. Developed over the course of five years, The Toolkit is a workshop format designed to help teams develop a sensitivity for how ethical values are connected to their own everyday work. Teams attend an introductory workshop, run by an ethics expert or trainer, where they learn about the intentions behind the Manifesto values and discuss what these might mean to their team and their work. After the introduction workshop, the teams name one member as their Ethics Guardian and receive materials so they can run smaller versions of the workshop for each project.

17.3.2 Research arrangements

The first author collaborated with two teams in the innovation department at The City. Both teams work on data driven solutions to challenges faced by the municipality, focusing on smart city projects, which utilize sensors and other digital data collection methods to provide services for both civil servants and citizens. The first author leveraged their prior expertise as a designer to collaborate with both teams on one specific product (redacted for anonymity) within the bigger Amsterdam4All project, which aims to make The City more accessible for people with mobility impairments. This work provided the background for the first author to be integrated into the teams and to learn about the daily integration of ethical initiatives into the work practices. The teams considered themselves to be responsible for showcasing what digital technologies and AI can do for the city. Following the release of ChatGPT, AI has received much attention within the municipality and teams working with the technology were facing lots of interest from other departments. The team therefore found itself in a position where it had to demonstrate possibilities and opportunities that were both enticing enough for other departments to make the effort of putting resources into adopting these new technologies, and feasible on a technical level.

17.3.3 Ethnographic engagement

The study was conducted both in person and online. The first author worked with the teams for 5 weeks in person and attended meetings and sessions online for an additional three months before and after the in person-period. During this time, they conducted interviews with all team members, attended regular team sessions such as stand ups, sprint planning, team retrospectives and task refinement sessions, and joined additional team sessions scheduled as needed. Even though the first author was on site for a specific project, interviews and meetings covered a broad range of past and ongoing projects in the city, that the teams were involved in. The first author also attended joint meetings between the team and other collaborators, and interviewed other employees of the city, such as managers or close collaborators. In total, the first author conducted 13 formal interviews, attended 30 team meetings, and 15 other non-team specific meeting sessions. Interviews were recorded, and meetings were documented through extensive note taking.

The first author embedded themselves into the ongoing project work amongst the two teams, and followed the practices and negotiations that are part of everyday work with the ethical materials. They were able to observe how the team members talk about and perceive their work and its ethical dimensions, as well as how they integrate the practical tasks into their everyday work, negotiate required resources and requirements, and schedule, scope and define resulting work tasks. This allowed the first author to experience how each team defines itself in relation to the rest of the organization, how they see their internal and external responsibilities, and their team purpose. The first author observed the affective impact of the engagement with ethical challenges of data driven design work, along with the impact of social relations on such affective experiences.

17.3.4 Data collection and analysis

The first author is experienced and trained in qualitative research methods through previous research projects and collaborations and has professional experience in industry through a full-time leadership position through which they could relate to and evaluate stories shared by the participants. Throughout the stay, they kept a diary for daily note taking and reflections. The research team maintained a weekly correspondence where the first author sent regular ethnographic reports to the rest of the team, engaging in joint discussion and reflection over email to facilitate ethnographic theorization (Cerwonka and Malkki, 2008). The ethnographic notes were transcribed, and first open coded, followed by coding for themes of interest and finally coding for language used to express emotional experiences and relational configurations. In addition to mapping the logistical challenges, we paid attention to the feelings emerging through these reports, and the emotional layers that shined through in the team members communications and actions.

In what follows below, we present different experiences of affective discomfort and the ways in which the practitioners navigated or avoided this through vignettes. These vignettes illustrate the intricate relational dynamics that ultimately contribute to or hinder the continuous engagement with the ethical tensions. To make our examples easier to follow we focus our vignettes on the team that had more experience of the materials although the conclusions illustrate our broader findings. The vignettes are written from the first perspective, as experienced by the first author.

17.4 Findings

Ethics toolkits and similar materials to operationalize ethics are often expected to function like any other organizational process, while the surrounding power structures and required relational labor is overlooked (Wong *et al.*, 2023), and are built on the assumption that teaching teams reflection and increasing their ethical awareness will result in positive outcomes (Vanhée and Borit, 2022). While much research points to how overlooking emotional aspects can explain why materials meet little adoption or acceptance (Frauenberger *et al.*, 2017; Madaio *et al.*, 2020; Metcalf *et al.*, 2019), it is also important to consider the affective impact on teams when they fully embrace the intentions and expectations of ethics initiatives, to ensure sustainable and successful integration. Engaging with ethics requires knowing and challenging one's own positionality (Drage *et al.*, 2024) and accepting a state of uncertainty and imperfection (Shklovski and Némethy, 2023). As the teams in our study engaged more deeply with The Manifesto and The Toolkit, they began to develop a sensitivity for the intertwined ethical dimensions of their work beyond its technical aspects. This resulted in conversations in which the team reflected upon and questioning of their work practices, which were noticeably laced through with affective language. Yet the affective experience of developing stronger ethical awareness in everyday settings is an often overlooked consideration to understanding how ethics tools may affect the work practice.

Our observations show that the affective experience of ethics work is shaped by circumstances depending on different relational configurations, similar to the notion of layers of micro, meso and macro concerns observed by Pillai *et al.* (2022). Yet where Pillai *et al.* focus on how broadly concerns may apply, we focus on the relational context of these concerns within organizational structures: within team, beyond team, and beyond organization.

These different relational configurations shape how team members experience the affective impact of their ethical awareness, and how they can respond to it. They also present varying demands on the team members on how to communicate and navigate the organizational structures they find themselves in. Within team describes the relational labor the team engages in to support each other in making the practice of The Toolkit possible. Beyond team describes the tensions the team experience in relation to their position within the organization, the expectations that are put on them by others, and the responsibilities they take on towards other teams. Beyond organization describes the tensions and concerns that stretch outside of their organization and touches on the relationship of the team as civil servants with the citizens of The City. For each configuration we first describe the framing and relevance, then provide two vignettes that describe particular situations that the first author observed or was part of (they are hence written in first person voice) and follow this with unpacking the vignettes. To provide a better overview of the people involved in the team, we start with introducing the team members and their roles (names changed):

Sarah is an AI researcher and has been with the team for 5 years. She has taken over the team lead position in the last year. She remembers that some conversations around ethical responsibility were present before, but with The Toolkit, became a more recognized part of everyday work practices.

Paul is an AI specialist and has joined the team 3 years ago. He was in other technical teams before and has worked a lot with student projects and other explorational projects. He has always been interested in inclusion and diversity and pushed projects such as Amsterdam-4-All.

Nadine has been with the team as data scientist for 4 years, and joined mostly because of an interesting project fit, after having previously been part of other teams at the municipality. She perceives her team members as generally kind people and appreciates that they think deeply about things, but she also thinks people in the team struggle with finding their concrete role.

Eli is a junior data analyst. She joined the team a little over a year ago on invitation by Sarah, after doing her bachelor thesis project at the municipality with some people from the team. She thinks the Manifesto values are important but prefers to work on the technical things and focuses on developing.

Johan is the project manager in the team. He has worked at the municipality as a freelancer for a few years and has joined the team officially about a year ago. He would like for the team to be more outward facing, and active in connecting and presenting their work and expertise to other teams and departments.

17.4.1 Within team

The team itself is the first space in which tech workers explore and potentially confront ethical dilemmas of their work. Following the initial educational workshop, the toolkit provides a smaller format of the workshop which can be run by the teams internally during each project they kick off. This means that if the teams are willing and open to continuing to integrate the toolkit into their own practice moving forward, the team itself provides the social frame for negotiation, confrontation and decision making around ethical dimensions of the projects discussed. The team therefore presents the smallest and most concentrated social space within which the affective experience of engaging with ethical dilemmas unfold. We therefore start with the dynamics amongst the team members themselves as first relational configuration through which to illustrate the emotional dimensions of ethical considerations in practice.

* * *

Vignette 1: Doing it wrong
It is my first week with the team. I am walking to an organizational meeting from one office to another with Paul. He is busy with various projects, but very invested in chatting with me, "to trade some gossip". He tells me about a lot of the prior projects they have worked on, and a lot of the concerns he has about them since starting to work with The Toolkit, often jumping between personal feelings and organizational struggles. When I explain that I still need to figure out where to spend my time during my stay, he responds excitedly "Oh can that be with us? I am really interested in hearing what you think after spending time with us! You can tell us what we are doing wrong!".

It is my last day with the team. I am having a closing conversation with Sarah. I share that I have already set a date with Johan in a few weeks so that I can share some insights from my visit. He was very eager to make sure that I report back to them so that they can learn from my visit. "Yeah, I think that is great. I mean you saw some of our difficulties" Sarah says. She gives examples such as running workshops too late, and not being in the habit of doing them first thing when starting a new project. "It would be great to hear what we can improve. I am sure there is so much we are not doing great at the moment."

* * *

Vignette 2: Value keeper

Nadine is sorting the outcomes of one of the ethical workshops into the tickets backlog of the team, while we chat. After the very first, big ethical workshop with an external facilitator, the team has decided that she will take on the role of Value keeper. She works through the tasks swiftly as she talks me through who she assigns to which ticket. Some tasks have very concrete to dos, while others are relatively undefined. In those cases, Nadine takes the freedom to fill in what she remembers from the prior discussions around why this task was deemed important and make judgements as to how they should be formulated on the ticket. Sometimes she splits tasks up into multiple tickets, if she thinks the chance is higher that they will get done that way. She feels that the process works pretty well, and the team trusts each other with being diligent in fulfilling the tasks to their best abilities. Others in the team appreciate the way she translates the abstract discussions during the workshops into concrete and actual tasks for them to act on. Despite doing a good job by all accounts, Nadine has issues with her role. “I really don’t like the term value keeper. It feels heavy. I am the kind of person when I have responsibility, I take it very seriously. So if you call me value keeper, it is just too much on my shoulder.”

* * *

4.1.1 Observation 1: Self-evaluation is stressful

Since the workshops and the work with The Manifesto are a self-organized activity, and something that no one else evaluates them on, the commitment to consistently scrutinize projects through the lens of the values asks of the team to challenge themselves. While the team considers this work as the morally right thing to do and appreciates having concrete tools at hand to engage with the ethical dilemmas of their work, they also remain unsure about the quality of their efforts. The team is left to their own devices to evaluate whether their efforts are “good enough”, tying the perception of their performance to their individual moral compasses. This leaves them with the moral responsibility of checking on each other and negotiating this moral metric amongst each other, potentially facing emotional stress through the uncomfortable feeling of moral uncertainty.

At the same time, the team is very open about their insecurity, looking for guidance from an outsider, and strive towards improving and developing. This mindset shows two things: first, the team is second-guessing their own efforts of ethical engagement with their work, but they are secure enough to be open about it and confident in their investment and the overall approach of these effort; second, they suspect that things are supposed to work or feel differently, easier, less burdensome, if they were doing it ’right.’ The team embraces the integration of The Toolkit into their processes, but doubts that they are living up to the expected impact this kind of effort is supposed to produce. This is potentially caused by an unrealistic expectation that The Toolkit workshops are a way to clarify the right thing to do.

4.1.2 Observation 2: Individualized responsibility is heavy
While the team members embrace the tools and experience support and openness within the team when facing doubts and insecurities, the affective experience of integrating the ethical work into their everyday processes nevertheless creates affective strain. The heaviness that Nadine notes is a byproduct of the individualized responsibility for morality that she takes on in the role of the value keeper. Even though she is personally invested in the ethical quality of her work, and cares about these dilemmas a lot, being put into a role of responsibility for moral conduct comes with the emotionally charged labor of moral judgement. She works to create explicit ties between the values and concrete tasks that she can put on the backlog for the next team sprint as the best way to integrate the responsibility as shared, "normal" work for the entire team, rather than special efforts. By integrating the considerations into existing flows and vocabulary, the work becomes more compartmentalized and less charged. However, the general nature of ethical questions is tricky to fit into the concise boxes of agile task management, especially if project requirements haven't been clearly defined by upper management with regard to the Manifesto values, and dilemmas haven't clearly been named.

17.4.2 Beyond Team

The team is embedded in the innovation department of the municipality, and hence part of a broader organization, within which they collaborate and communicate with a variety of other teams and stakeholders. While the team partially picks their own projects to work on, they also receive projects to develop from higher management, and collaborate with other teams in the municipality on their project ideas. Two aspects are particularly important in this relational configuration: first, the team builds projects not only for citizens but also for internal uses and other departments; second, the team has managers who are not necessarily involved in everyday work structures. While the team needs to internally navigate how to address ethical dilemmas and the additional, potentially uncomfortable work amongst each other, they are also confronted with pressures, expectations and tensions by other departments and existing organizational structures, which do not engage in the same work of reflection through The Workshops.

* * *

Vignette 3: The tech hype and fear of missing out
With the rise of technologies like ChatGPT, the team is experiencing more interest from other sections in the department, who are curious about using similar services for their work. "It seems that even though there's all this fear mongering, I think when people want to use AI for their work, they just want the benefits and they don't

want to think about it right," says Sarah, and she goes on to describe that people get many impressions from the media and then turn to their team with questions and ideas. She does not feel sure that the municipality has actually put proper consideration into whether they want to support this technology as a municipality or is responding to a demand that has been creating because of the release of ChatGPT and feel pressured to provide their own solution.

Nadine has a similar opinion, when I ask her about the general position of the team in the organization, and how they relate to collaborating partners. "Yeah, one of the challenges I do think is, like, the tech hype, and that there is like, oh, there's this interesting new technology, we have to do something with it. So people are sort of expecting something, but really not clear about what they're expecting. So you're in some weird middle ground where you have on one hand really a lot of freedom, but at the same time, not that much freedom because so many people are watching you."

* * *

Vignette 4: Experts with vague gut feelings

The team has received the task from Martin, their upper manager, to work on a version of ChatGPT for the municipality "so we can put that out there and make informed expert recommendations". However, the team keeps running into confusion about the project goals. The question of "who are we doing this for and why are we building this?" comes up in three meetings: the ethics workshop, the backlog refinement, and the sprint planning. Sarah struggles with pushing the team towards getting the project going, and accounting for the many open questions in this project. She regularly brings up the responsibility of the team as civil servants as well as technical experts "We, as a team, have the same responsibility [as all civil servants to raise issues and concerns] and maybe even a stronger responsibility when it comes to AI." During the backlog meeting Johan points out that, "they [the steering group] also have ethical people there, probably legal, so that is not our responsibility". To which Sarah agrees, but is still concerned, particular about the more uncertain questions the team is having about the initiative. "The less clear ones are, of course, the more vague ethical feelings that you might have about something where, you know, you could somehow frame it in a way that meets all the official regulations, but still, your gut feeling says that it's not a great idea." Eventually Sarah agrees hesitantly with Johan: "In the end, it's going to be a sort of political choice that a manager at some point needs to sign off on."

* * *

Observation 3: Informal gate keepers

The exposure to The Manifesto and The Toolkit has helped the team to become more ethically aware of the broader impact of their work. Beyond providing technical expertise on whether something is technically feasible or not, the team also wants to make sure that they take into account the ethical impacts of potential future services and feel the need to step beyond their position as expert producers and cover specific ethical viewpoints as well. By taking on the responsibility created by The Manifesto as something they should embody within their projects, the team puts the potential ethical costs of projects also on their plate to figure out and attend to. This requires of the team to set boundaries around cost and benefit beyond technical expertise. In addition, the team takes on the role as an informal gate keeper, where they feel responsible to consider the ethical consequences of potential projects, without necessarily having the mandate to actually curb them – which is generally associated with legal teams, who act as gatekeepers on legislative grounds. This often results in frustration about projects they are assigned.

Observation 4: Mismatched expectations

The Workshops raise ethical dilemmas, but these dilemmas are rarely a clear-cut trade-off between benefits and costs, or address situations where a benefit might just as well be a cost and vice versa. In these cases, the team is unable to resolve dilemmas through any kind of workshop or reflection, but instead needs a political decision, which is beyond the team's technical mandate. The willingness of the team to reflect beyond the technical dimensions of their potential projects confronts them with the limitations of their own roles and responsibilities, creating frustration with the lack of direction. As the technical team, they are the experts that the municipality comes to for consultations about feasibility and testing, but the team is already further in their own ethical reflection and thus run up against the limited understanding of the ethical concerns beyond their own group. They must manage not only their own discomfort around the uncertainty of their own decisions, but also need to navigate the emotional currents of the teams and management around them, such as reactive projects based on external pressures, hype, and fear of missing out.

Observation 5: The question of power

The consideration of doing something that is not welcomed outside of the team, but instead runs counter of what is expected, confronts the team with their own positioning within the organization. While the engagement with the toolkits results in a stronger ethical awareness as intended - or what Sarah refers to as vague gut feelings - the surrounding organizational structures are not built to attend to vague feelings when trying to estimate ethical costs and utility benefits from technical solutions. When the lines between regulation and ethical concerns, or risk and benefit are not obvious, the team will not have the last say in what will be done. While the team is supported through the framing of the tools to open conversations amongst each other, and potentially even with other teams and management, the resulting way forward is grounded in political decisions that are out of the team's locus of control. In other words, the tools give the team the power to reflect on the ethical dimensions of their work, but only limited power over the consequences of such reflection and the mandate to address them (Seberger *et al.*, 2021). Organizational structures are thereby both a source for ethical tensions and a limiting factor for resolution. This is a consequence of "ethics" being isolated as part of the development and design process, and the existing materials such as The Toolkit do not include channels of escalating such dilemmas towards higher levels of management. Toolkits, such as The Toolkit, thus do not account for the fact that teams will inevitably run into the limits of their own organizational power.

17.4.3 Beyond organization

Finally, the municipality is part of the general city, together with citizens, businesses, and other organizations. As part of the municipality, the team must grapple with tensions that arise between their own projects and intentions and the interests of the citizens they are serving.

* * *

Vignette 5: On being transparent and feeling naked
One of the first initiatives the team mentions to me is of the Cities algorithm bank, where technical teams are supposed to commit their code, to ensure that all models are openly viewable. "Commit code to algorithm register," "make sure data is openly available," and "make sure to ensure privacy when making data public" pop up on many Miro boards during ethical workshops for multiple projects. During one ethical workshop, Paul asks whether they ever got around to uploading the data from the prior project. "I think we got a bit scared" responds Nadine laughingly, but also a bit apologetic. Often, the task of uploading the code lingers, even though it is one of the most straightforward tasks in the backlog. When I ask Nadine about it she responds "Yeah, I think only the thing... putting our code open source is always taking a long time and it's also just people being sort of ....you feel kind of naked publishing your code, right. So I really believe in everyone should do it. But once we started working on it, we start to get a little bit nervous and perfectionist and yeah, so that's always a tricky one."

* * *

Vignette 6: Innovation and the Status Quo
After having worked on two projects that address the quality of sidewalks in The City, Eli has started to see the infrastructures of the city differently. She now notices bikes that are being parked wrongly and block the sidewalks or other obstacles in the streets. She makes sure that she doesn't park her bike in the way of a potential wheelchair path and sometimes she even moves others when she can. During a workshop, she discusses with the others how to draw attention from officials to the issues of lacking infrastructure and challenges for mobility impaired citizens. The group agrees that to match the ideals of The Manifesto, broader change of

infrastructure is needed, more than they can do with an app. During the ethics workshop the next day Nadine reflects upon how the realization that building an app that supports people with mobility impairments might relieve pressure off other departments to design the city more inclusively in the first place, really hit her as being the opposite of what the team actually would like to achieve: “Like you have discussions and for some things, you think okay, yeah, we could be a bit more transparent, that’d be cool. But this one like was really like hitting me in the face like, okay, I think we’re doing . . . are we not really consolidating the situation instead of changing it?”

* * *

Observation 6 – Decisions are hard, actions are harder

The team members understand themselves as civic servants, and The Manifesto itself postulates its values in relation to citizens of The City, such as stating that data usage should benefit everybody, that models should be transparent for everybody to understand, etc. As such, they have a strong perception of how they ought to act, and do a good job, of formulating that during the action of matching values to tasks. However, while the workshops give the team the tools to decide on ways forward – i.e., the tasks they want to take – the actual discomfort around these tasks often surfaces outside of the workshops, when it comes to actually performing the tasks, in particular when it comes to actions that relate to stakeholder outside of the organization. Here, the team most strongly encounters the tension not of holding others accountable or confronting them with their ethical dilemmas, as they do in the beyond-team vignettes, but to hold themselves accountable towards a faceless, anonymous group of people they feel beholden to, which confronts them with feeling exposed, naked, and conscious of what the quality of their work communicates to outsiders.

Observation 7 – Reflections cause more questions, not answers

The team struggles with the problem that the systems they encounter in their projects are built into an environment with already established structures that pull towards a status quo. This has two consequences for the team: first, these structures are not in line with the values in the projects and thus can easily subvert them; second, the team can only make progress from the given starting point, but tends to judge itself from an ideal interpretation of the values. Any fundamental change in infrastructure is outside of the mandate or jurisdiction of the team, and thus out of their reach. Even when attempting to live up to the idealistic values, the team also starts to recognize that sometimes, tech solutions are just band aides to bigger issues, and all their efforts to improve the ethical scoring of their production process – by making it more inclusive, transparent etc. – might not create the end result they would prefer, simply given the fact that their potential solutions are tools and not systemic changes. Reflecting upon the far-reaching sociopolitical consequences of technical interventions is a core aspect of what ethical toolkits encourage developers to do. However, deep reflections on far reaching consequences facilitated by the tools often spark more questions and realizing that the ethics work that they are expected to do is more treating symptoms rather than the disease. Rarely is the affective outcome thematized in these approaches, such as being hit in the face with the realization that high quality, well-meant technical interventions might contribute to consolidating existing exclusion structures, leads to rather existential questions of the contributions of technical teams in general.

17.5 Discussion

Our research shows that the presence of The Manifesto and The Toolkit affect the ways in which the teams considered their work, most evident in their recognition of the broader social and political impact of their work and the reflections on their own position and responsibility. Researchers have long argued that for ethics interventions to be successful there must be a combination of emotionally invested and open-minded practitioners willing to engage and reflect (Frauenberger *et al.*, 2017; Metcalf *et al.*, 2019), and tools developed with existing work structures and practical requirements in mind (Madaio *et al.*, 2020). In our study, the combination of emotionally and responsibly invested team and actionable toolkit provided a pretty good example of an ideal case, where human factors and resources fit well together.

However despite these “ideal” circumstances, the teams in our study struggle with constantly feeling that they were not quite “getting it right” in their efforts to sufficiently think through ethical implications of their work (Vignette 1, 2, 4; Observation 1, 2). The vignettes across the relational configurations of within team, beyond team, and beyond organization, show that even well-developed interventions have significant and uncomfortable emotional effects on teams who are ethically aware and engaged yet find themselves faced with many challenges and few certainties, even when responding positively to these interventions. Worse, neither the tools that are intended to support ethical reflection nor the organization that has put significant resources into the development and implementation of these tools, are able to provide support for the feelings of vulnerability, heaviness and frustration that seem to be endemic to serious ethical reflection that is part of technical practice.

17.5.1 Beyond reflection: Ethics in practice as emotional labor

One main tenet of toolkits such as The Toolkit is to provide scaffolding for deeper reflection on the social, political, ecological, and economic impacts of tech work beyond the primary goals of utility and efficiency. The reflections sparked by The Toolkit workshops created ethical awareness that caused the technologists to see new aspects of their work and give weight to moral considerations in their choices

(Observation 1, 6, 7). This phenomenon is well observed and discussed in moral psychology from biological, psychological and socio-cultural viewpoints under the concept of moral awareness (Jordan, 2009; Reynolds *et al.*, 2012). While many ethics studies and toolkits in the tech field refer to moral (or ethical) awareness or sensitivity as one of their intended goals (Tangible, 2024; IDEO, 2019; ARTEFACT, 2024), both the tools and the research tends to overlook the psychological aspects of moral awareness. Studies in moral psychology have for example shown that while higher m¬oral awareness results in more self-reflective behavior, it will also make people more self-conscious and uncertain of their actions (Reynolds and Miller, 2015). In our findings we observed two tensions that emerged due to the notion that deeper reflection will provide better outcomes without considering consequences of moral awareness to the individual or the surrounding organization.

First, ethics tools are a moral comment on the existing practices (Observation 1, Vignette 1). Introducing new ethics-oriented practices poses the question on whether work was morally lacking before. Will ethical reflection improve the resulting work and products because they receive an ethical upgrade, (Observation 1, 2) or did the people have shortcomings before and are better trained now, in which case the question would be better trained for what? Asking technologists to reflect upon their work demands them to face and recognize potential wrongdoings in the past and to respond accordingly as a consequence of reflection (Reynolds and Miller, 2015). Integrating such recognition into their future work, demands emotional labor of the team members individually and collectively, working out the boundaries what they, as a team, might deem morally acceptable, in order to accommodate a shifting perception of their roles and responsibilities (Observation 4, 7; Vignette 4, 5).

Second, reflecting on and recognizing the multi-layered character of tech work is always going to lead to a position of "what now?" where people must make decisions, take stances, and coordinate actions (Observation 1, 3, 7; Vignette 3, 5). The ethical toolkits offer little support for concrete action, but create new social and emotional "costs" – trade-offs the teams encountered in their everyday work to fit the outcomes of the ethics workshops into their practices. Such costs can be logistical and procedural, such as spending more time and money on a project to accommodate additional work or changed features (Observation 6, 2; Vignette 2, 4). In case of project course corrections, such costs can be political, in that the team needs to put extra efforts into negotiating and convincing across layers of hierarchy (Observation 3,4 , 5; Vignette 3, 4). If neither project modifications nor course corrections are possible or feasible, the team needs to compromise on their own ethical standards (Observation 3), leading to emotional costs in the sense of compromising personal values.

These costs lead to what moral psychology names "moral stress:" technologists embracing increased moral awareness will experience feelings of discomfort, vulnerability, and exposure (Vignette 5), which are part and parcel of the inherent doubts and uncertainties of ethical decision making (Shklovski and Némethy, 2023). Moral stress can occur when a person realizes that they need to make ethical decisions but there is a mismatch between expectations and possible actions (Reynolds *et al.*, 2012). This can occur when a person makes a moral decision to act but is prevented from following through due to circumstances beyond their control, or when results of reflection conflict with existing practices (Reynolds *et al.*, 2012). Feelings of vulnerability then are more than "an active ethical stance" as Popova and colleagues (2022) argue . Rather, they are endemic to making ethical considerations explicit in technical processes and practices and need to be supported because, as research suggests, ignoring moral stress can lead to resignation (Corley *et al.*, 2001; Imbulana *et al.*, 2021).

17.5.2 The feeling of "Not getting it right"

While no ethics interventions claim to provide solutions to ethical dilemmas, they typically position themselves as guidance or support in making decisions, thus easing the burden and improving outcomes (Chivukula *et al.*, 2022; Morley *et al.*, 2020). These are often situated in a culture coined by solutionist thinking, where problems are inherently coupled with solutions, even if the problem is inherently irresolvable (Cunningham *et al.*, 2023; Sicart and Shklovski, 2020). When it comes to value sets and ethical toolkits, such coupling creates expectations that dealing with and addressing ethical dilemmas through the provided exercises is a skill to be trained, becoming easier with time and resulting in 'better' outcomes. These expectations fall apart when aspirational values and idealistic resolutions collide with real life working conditions, and the exercises fail to provide paths of action (Vignette 4, 6). Instead, the deeper engagement with ethics and the entanglement of articulated values with work practices surfaces ethical dilemmas that can be difficult or impossible to resolve, either because they are inherently in conflict with each other or because they confront the limitations instilled through the existing status quo (Observation 3, 4, 6). Such unresolvable dilemmas cause moral stress but there is typically no outlet for resolving it (Reynolds *et al.*, 2012). Facing dilemmas and taking an active stance – whichever stance that might be – makes the teams unavoidably vulnerable to being challenged (Popova *et al.*, 2023). Yet there are typically few organizational infrastructures available to support such vulnerability.

The teams in our study were predictably unable to make everyone happy and needed to navigate compromises and sub-optimal process decisions, while at the same time wondering where they fall short or what they are doing wrong (Vignette 1, 2; Observation 7). On some level the teams were aware that they will inevitably fail to some extent and will never be able to fully reach the idealistic vision portrayed by The Manifesto. As the ethics work was located within the teams without external intervention, they struggled to recognize what would be 'good enough' or 'right enough' while The Workshops inevitably created more questions than answers.

Increased moral awareness turns out to look and feel very different than what the teams expected and what The Toolkit designers lead them to believe. Rather than finding clarity and certainty, feeling more assured in the goodness of their work and capable of delivering morally good projects, they experienced more uncertainty and less confidence in their projects. Instead of finding clarity, they identified and negotiated increasingly difficult tradeoffs, and instead of having certainty, they had to face the discomfort and heaviness of their responsibilities (Vignette 2, 4, 6; Observation 2, 4 5). Such affective experience of moral stress (Reynolds *et al.*, 2012) goes counter to the general expectations of what tools, pipelines and work processes are expected to do in a technology innovation context. Acknowledging this persistent failing, but still doing better, and getting at least closer to a potentially more desirable and morally aligned outcome, requires the team to continuously stay with their discomfort and cope with the moral stress they experience, however they manage. In our study, the teams persisted in doing the hard thing – such as finding compromises, adding extra tasks to their project that they feel lie outside their expertise and generally navigate the emerging spaces of uncertainty together (Shklovski and Némethy, 2023). Yet this was only possible because the team dynamics functioned very well and allowed them to cope with uncertainty and discomfort. Counter-intuitively for the team, the positive effects of increased moral awareness is the capacity to reflect as a team, to continuously question and reconsider while not getting stuck or paralyzed from potential implications of design decisions (Frauenberger *et al.*, 2017) (Vignette 2, 5).

17.5.3 Responsibilities of organizations: beyond technical practice

Ethical practices that are anchored to individuals instead of organizations, fall short of impact and put the individuals into difficult and draining positions (Raji *et al.*, 2021; Widder and Nafus, 2023) (Vignette 2, 4, 5; Observation 2, 5). The Toolkit localized "ethics" in technology within very confined team structures of the people producing the code or making direct design decisions. As a result, where the teams in our study clearly increased their morale awareness this was not happening throughout the organization. While worrying about high-level concerns of fairness and transparency, The Toolkit did not account for the organizational infrastructures of power where these teams are embedded (Observation 5, 7). In this context, the teams needed to negotiate the demands of managers and other decision-makers, attempting to influence broader organizational processes and then eventually giving up and moving on. Such experiences were another source of moral stress that the teams had to navigate (Vignettes 4, 5; Observation 3, 4).

The teams often described "doing the ethics work" as difficult, hard, and a "heavy" responsibility both for the individual contributors and for the team to navigate,. Rather than an indicator of shortcomings or failing work processes, these discomforts were also indicators of the team approaching the required commitment with emotional investment and openness, hence, doing it "right". Yet aside from conversations with the first author as an accidental available 'expert' they had no way of assuaging these discomforts. Instead of expecting optimistic attitudes, organizations need to recognize that if this type of work is not hard, the teams are likely not engaging in it fully (Girard, 2011). This, unfortunately, is often in contrast with common business practices where additional questions or uncertainties rock the boat both within teams and outside of them (Girard, 2011; Irani, 2019).

For ethics interventions to succeed, teams need emotional scaffolding in addition to logistical resources. Organizations must internalize that ethics considerations are tricky, and projects will become more openly complicated when surfacing ethical tensions. In order for employees to take on these tensions, owning up to potential mistakes through the necessary self-critical reflection, they must be able to trust that they will not face the danger of being scapegoated, criticized or put down for decisions that do not align with internal organizational prospects, such as delivery timelines or financial targets (Drage *et al.*, 2024). Working through ethical tensions therefore needs to be a full-scale organizational commitment, where moral stress is recognized as legitimate through supportive structures. Even if reflection work and ethical sensitivity can be developed from the bottom up, the management of the emotional cost needs to be carried and accounted for from the top. Otherwise, were employees are left to deal with the moral stress on their own without sufficient power to create meaningful change, teams run the risk of surrendering to the pressures of moral stress, potentially choosing ignorance and resignation as coping mechanisms (Corley *et al.*, 2001).

17.6 Conclusion

In this paper, we discuss the affective experience of committing to integration of ethics driven interventions in technical work practices. Even when toolkits are designed with logistical necessities in mind, and technologists are emotionally invested and motivated to "do the right thing", ethics interventions require emotional labor and produce moral stress, in the form of vulnerability, discomfort and other difficult emotions. It seems however that interventions such as ethical toolkits consider their recipients as disembodied, impartial actors, who can engage with and work through the expected process of reflection, identity work and required action without any emotional reaction to the impact this work has on shifts of perspective and awareness of positionality and responsibility. We argue that a successful and sustainable implementation of ethics initiatives requires accounting for inevitable moral stress (Reynolds *et al.*, 2012). Going beyond typical organizational resources such as project management and communication structures, organizations must provide emotionally supportive scaffolding that acknowledges the extra costs of moral awareness and moral stress (Corley *et al.*, 2001; Reynolds and Miller, 2015). Beyond the guidance on how to "do" ethics, teams require support in learning how to acknowledge and navigate the affective experiences they encounter, and receive reassurance that uncomfortable emotions are not necessarily a sign of bad conduct, but an inevitable part of the experience (Imbulana *et al.*, 2021).

Ethics interventions are also often located within technical teams, without any expectations of shifts in organizational practice. While organizations might be willing to support their technical teams by providing extra time and resources for integrating new ethics-related tools and processes, few recognize the necessity for broader organizational change. We show that teams that embrace moral awareness and attempt to influence organizational decisions accordingly often encounter barriers in the form of managerial hierarchies and internal politics. Such barriers can lead to frustration, disengagement and stunted impact of any intervention that aims to develop moral awareness without political reach. We see further promising directions for this research in deepening the understanding of how ethics initiatives can address the socio-political dimensions attached to these organizational practices and hope to motivate other researchers to follow a path that recognizes practitioners as emotional beings with affective needs when grappling with responsibilities, ethical tensions and uncertain consequences.

Bibliography 18

2038 The New Serenity (2022). en.

33A (2022). *AI Design Sprint™*. 33A. URL: `https://www.33a.ai/ai-design-sprint` (visited on Apr. 26, 2024).

Aarts, Tessa, Linas K. Gabrielaitis, Lianne C. de Jong, Renee Noortman, Emma M. van Zoelen, Sophia Kotea, Silvia Cazacu, Lesley L. Lock, and Panos Markopoulos (July 2020). „Design Card Sets: Systematic Literature Survey and Card Sorting Study". In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. DIS '20. New York, NY, USA: Association for Computing Machinery, pp. 419–428.

Abdi, Noura, Tess Despres, Ruba Abu-Salma, and Julia Bernd (2022). „In-Home Smart Devices: Quantifying Bystander Privacy Experiences and Social Norms in Different Situations". English. In: *Annual Symposium on Applications of Contextual Integrity (CI Symposium)*.

Ackermann, Rebecka (Sept. 2, 2023). *Design thinking was supposed to fix the world. Where did it go wrong?* MIT Technology Review. URL: `https://www.technologyreview.com/2023/02/09/1067821/design-thinking-retrospective-what-went-wrong/` (visited on Oct. 30, 2023).

AI, Google (2023). *Why we focus on AI*. Google AI. URL: `https://ai.google/why-ai/` (visited on Aug. 31, 2023).

Akrich, Madeleine (1992). „The De-scription of Technical Objects". In: *Shaping Technology/Building Society. Studies in Sociotechnical Change*. Shaping Technology/Building Society. Studies in Sociotechnical Change. MIT Press, p. 205.

Albrechtslund, Anders (Mar. 1, 2007). „Ethics and technology design". In: *Ethics and Information Technology* 9.1, pp. 63–72.

Alhammad, Manal M. and Ana M. Moreno (Oct. 17, 2022). „Integrating user experience into Agile: an experience report on lean UX and Scrum". In: *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Software Engineering Education and Training*. ICSE-SEET '22. New York, NY, USA: Association for Computing Machinery, pp. 146–157.

Ali, Safinah, Vishesh Kumar, and Cynthia Breazeal (Sept. 6, 2023). „AI Audit: A Card Game to Reflect on Everyday AI Systems". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.13. Number: 13, pp. 15981–15989.

Alkhatib, Ali (May 7, 2021). „To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes". In: *Proceedings of the 2021 CHI Conference on Hu-*

man Factors in Computing Systems. CHI ’21. New York, NY, USA: Association for Computing Machinery, pp. 1–9.

Allen, Philip van (Oct. 25, 2018). „Prototyping ways of prototyping AI“. In: *Interactions* 25.6, pp. 46–51.

Ames, Morgan G. (Aug. 17, 2015). „Charismatic technology“. In: *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives*. CA ’15. Aarhus N: Aarhus University Press, pp. 109–120.

Ames, Morgan G. (Nov. 19, 2019). *The Charisma Machine: The Life, Death, and Legacy of One Laptop per Child*. Google-Books-ID: v4y5DwAAQBAJ. MIT Press. 323 pp.

Ames, Morgan G., Janet Go, Joseph ’Jofish’ Kaye, and Mirjana Spasojevic (Mar. 19, 2011). „Understanding technology choices and values through social class“. In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. CSCW ’11. New York, NY, USA: Association for Computing Machinery, pp. 55–64.

Anderson, Ken, Maria Bezaitis, Carl Disalvo, and Susan Faulkner (2019). „A.I. Among Us: Agency in a World of Cameras and Recognition Systems“. In: *Ethnographic Praxis in Industry Conference Proceedings* 2019.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/155 8918.2019.01264, pp. 38–64.

Andrus, McKane, Elena Spitzer, Jeffrey Brown, and Alice Xiang (Mar. 1, 2021). „What We Can’t Measure, We Can’t Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness“. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. New York, NY, USA: Association for Computing Machinery, pp. 249–260.

ARTEFACT (2024). *The Tarot Cards Of Tech*. URL: `https://tarotcardsoftech.artefactgroup.com/` (visited on Apr. 26, 2024).

Auger, James (Mar. 1, 2013). „Speculative design: crafting the speculation“. In: *Digital Creativity* 24.1. Publisher: Routledge _eprint: https://doi.org/10.1080/14626268.2013.767276, pp. 11–35.

Ayling, Jacqui and Adriane Chapman (Aug. 1, 2022). „Putting AI ethics to work: are the tools fit for purpose?“ In: *AI and Ethics* 2.3, pp. 405–429.

Ballard, Stephanie, Karen M. Chappell, and Kristen Kennedy (June 18, 2019). „Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology“. In: *Proceedings of the 2019 on Designing Interactive Systems Conference*. DIS ’19. New York, NY, USA: Association for Computing Machinery, pp. 421–433.

Ballo, Ingrid Foss (Sept. 1, 2015). „Imagining energy futures: Sociotechnical imaginaries of the future Smart Grid in Norway“. In: *Energy Research & Social Science*. Special Issue on Smart Grids and the Social Sciences 9, pp. 9–20.

Bannon, Liam (July 1, 2011). „Reimagining HCI: toward a more human-centered perspective“. In: *Interactions* 18.4, pp. 50–57.

Bardzell, Jeffrey and Shaowen Bardzell (Aug. 2015). „The user reconfigured: on subjectivities of information“. In: *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives*. CA ’15. Aarhus N: Aarhus University Press, pp. 133–144.

Bardzell, Shaowen (2010). „Feminist HCI: taking stock and outlining an agenda for design“. In: *Proceedings of the 28th international conference on Human factors in* *computing systems - CHI ’10*. the 28th international conference. Atlanta, Georgia, USA: ACM Press, p. 1301.

Bareis, Jascha and Christian Katzenbach (Sept. 1, 2022). „Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics“. In: *Science, Technology, & Human Values* 47.5. Publisher: SAGE Publications Inc, pp. 855–881.

Baumann, Karl, Benjamin Stokes, François Bar, and Ben Caldwell (June 26, 2017). „Infrastructures of the Imagination: Community Design for Speculative Urban Technologies“. In: *Proceedings of the 8th International Conference on Communities and Technologies*. C&amp;T ’17. New York, NY, USA: Association for Computing Machinery, pp. 266–269.

Baumer, Eric P. S. and Jed R. Brubaker (May 2, 2017). „Post-userism“. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17: CHI Conference on Human Factors in Computing Systems. Denver Colorado USA: ACM, pp. 6291–6303.

Beck, Kent L., Michael A. Beedle, A. V. Bennekum, *et al.* (2013). „Manifesto for Agile Software Development“. In.

Benjamin, Garfield (June 20, 2022). „#FuckTheAlgorithm: algorithmic imaginaries and political resistance“. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. New York, NY, USA: Association for Computing Machinery, pp. 46–57.

Bennett, Cynthia L. and Os Keyes (Mar. 2, 2020). „What is the point of fairness? disability, AI and the complexity of justice“. In: *SIGACCESS Access. Comput.* 125, 5:1.

Bernd, Julia, Ruba Abu-Salma, Junghyun Choy, and Alisa Frik (2022). „Balancing Power Dynamics in Smart Homes: Nannies’ Perspectives on How Cameras Reflect and Affect Relationships“. en. In: *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pp. 687–706.

Bevilacqua (Nov. 2017). *Tesla Car With Self-Driving Features Strikes and Kills UK Cyclist*. en-us. Section: News.

Bidwell, Nicola J. (June 13, 2016). „Decolonising HCI and interaction design discourse: some considerations in planning AfriCHI“. In: *XRDS* 22.4, pp. 22–27.

Bietti, Elettra (Jan. 27, 2020). „From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy“. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, pp. 210–219.

Bilal, Adil, Stephen Wingreen, and Ravishankar Sharma (Apr. 20, 2020). „Virtue Ethics as a Solution to the Privacy Paradox and Trust in Emerging Technologies“. In: *Proceedings of the 3rd International Conference on Information Science and Systems*. ICISS ’20. New York, NY, USA: Association for Computing Machinery, pp. 224–228.

Binns, Reuben, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt (Apr. 21, 2018). „’It’s Reducing a Human Being to a Percentage’: Perceptions of Justice in Algorithmic Decisions“. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. New York, NY, USA: Association for Computing Machinery, pp. 1–14.

Birhane, Abeba (Feb. 12, 2021). „Algorithmic injustice: a relational ethics approach“. In: *Patterns* 2.2, p. 100205.

Birhane, Abeba and Fred Cummins (Dec. 16, 2019). „Algorithmic Injustices: Towards a Relational Ethics“. In: *arXiv:1912.07376 [cs]*. arXiv: 1912.07376.

Birhane, Abeba, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy (June 20, 2022). „The Forgotten Margins of AI Ethics“. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’22. New York, NY, USA: Association for Computing Machinery, pp. 948–958.

Bjerknes, Gro, P. Ehn, M. Kyng, and K. Nygaard (Aug. 1, 1987). „Computers and Democracy: A Scandinavian Challenge“. In.

Blythe, Mark (May 2, 2017). „Research Fiction: Storytelling, Plot and Design“. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. New York, NY, USA: Association for Computing Machinery, pp. 5400–5411.

Boddington, Paula (2017). *Towards a Code of Ethics for Artificial Intelligence*. Artificial Intelligence: Foundations, Theory, and Algorithms. Cham: Springer International Publishing.

Bødker, Susanne (Aug. 2015). „Third-wave HCI, 10 years later—participation and sharing“. en. In: *Interactions* 22.5, pp. 24–31.

Boess, Stella and Floris Jansen (June 25, 2022). „Values arising from participatory inclusive design in a complex process“. In: *DRS Biennial Conference Series*.

Bond, Shannon (Oct. 6, 2020). „How Are Apple, Amazon, Facebook, Google Monopolies? House Report Counts The Ways“. In: *NPR*.

Bonnemains, Vincent, Claire Saurel, and Catherine Tessier (Mar. 1, 2018). „Embedded ethics: some technical and ethical challenges“. In: *Ethics and Inf. Technol.* 20.1, pp. 41–58.

Borgmann, Albert (1984). *Technology and the Character of Contemporary Life: A Philosophical Inquiry*. Google-Books-ID: qS3pJL_BcdkC. University of Chicago Press. 310 pp.

Borgmann, Albert (June 1993). *Crossing the Postmodern Divide*. Chicago, IL: University of Chicago Press. 182 pp.

Borning, Alan and Michael Muller (May 5, 2012). „Next steps for value sensitive design“. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’12. New York, NY, USA: Association for Computing Machinery, pp. 1125–1134.

Bowker, Geoffrey C. and Susan Leigh Star (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge, Massachusetts: The MIT Press.

Boyd, Karen L. and Katie Shilton (July 13, 2021). „Adapting Ethical Sensitivity as a Construct to Study Technology Design Teams“. In: *Proceedings of the ACM on Human-Computer Interaction* 5 (GROUP), 217:1–217:29.

Braithwaite, Phoebe (July 22, 2018). „Smart home tech is being turned into a tool for domestic abuse“. In: *Wired*. Section: tags.

Braun, Virginia and Victoria Clarke (2012). „Thematic analysis“. In: *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative,* *neuropsychological, and biological*. APA handbooks in psychology®. Washington, DC, US: American Psychological Association, pp. 57–71.

Brey, Philip (Dec. 1, 2000). „Disclosive computer ethics“. In: *ACM SIGCAS Computers and Society* 30.4, pp. 10–16.

Brey, Philip A. E. (Dec. 1, 2012). „Anticipating ethical issues in emerging IT“. In: *Ethics and Information Technology* 14.4, pp. 305–317.

Brundage, Miles, Shahar Avin, Jack Clark, *et al.* (Feb. 20, 2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. arXiv: 1802.07228[cs].

Burrows, Alison, David Coyle, and Rachael Gooberman-Hill (Mar. 2018). „Privacy, boundaries and smart homes for health: An ethnographic study“. en. In: *Health & Place* 50, pp. 112–118.

CA (Apr. 25, 2023). *Directive on Automated Decision-Making*. Treasury Board of Canada, Secretariat. Last Modified: 2023-04-25. URL: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592 (visited on July 14, 2024).

Caire, Patrice, Assaad Moawad, Vasilis Efthymiou, Antonis Bikakis, and Yves Le Traon (Jan. 2016). „Privacy challenges in Ambient Intelligence systems“. en. In: *Journal of Ambient Intelligence and Smart Environments* 8.6. Publisher: IOS Press, pp. 619–644.

Calvi, Alessandra and Dimitris Kotzinos (June 12, 2023). „Enhancing AI fairness through impact assessment in the European Union: a legal and computer science perspective“. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’23. New York, NY, USA: Association for Computing Machinery, pp. 1229–1245.

Calvo, Rafael A. and Dorian Peters (Nov. 28, 2014). *Positive Computing: Technology for Wellbeing and Human Potential*. The MIT Press.

Campolo, Alexander and Kate Crawford (Jan. 8, 2020). „Enchanted Determinism: Power without Responsibility in Artificial Intelligence“. In: *Engaging Science, Technology, and Society* 6, p. 1.

Card, Dallas and Noah A. Smith (May 8, 2020). „On Consequentialism and Fairness“. In: *Frontiers in Artificial Intelligence* 3, p. 34. arXiv: 2001.00329[cs,stat].

Carstensen, Martin B. and Vivien A. Schmidt (Mar. 15, 2016). „Power through, over and in ideas: conceptualizing ideational power in discursive institutionalism“. In: *Journal of European Public Policy* 23.3, pp. 318–337.

Cave, Stephen and Kanta Dihal (Feb. 2019). „Hopes and fears for intelligent machines in fiction and reality“. In: *Nature Machine Intelligence* 1.2. Number: 2 Publisher: Nature Publishing Group, pp. 74–78.

Cervantes, José-Antonio, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos (Apr. 1, 2020). „Artificial Moral Agents: A Survey of the Current Status“. In: *Science and Engineering Ethics* 26.2, pp. 501–532.

Cerwonka, Allaine and Liisa H. Malkki (Nov. 15, 2008). „Improvising Theory: Process and Temporality in Ethnographic Fieldwork“. In: *Improvising Theory*. University of Chicago Press.

Chamberlain, Stephanie, Helen Sharp, and Neil Maiden (2006). „Towards a Framework for Integrating Agile Development and User-Centred Design“. In: *Extreme*

Programming and Agile Processes in Software Engineering. Ed. by Pekka Abrahamsson, Michele Marchesi, and Giancarlo Succi. Berlin, Heidelberg: Springer, pp. 143–153.

Chang, Emily (2019). *Brotopia*. Portfolio. 336 pp.

Chivukula, Sai Shruthi (July 22, 2021). „Designing for Co-Creation to Engage Multiple Perspectives on Ethics in Technology Practice“. thesis. Purdue University Graduate School.

Chivukula, Shruthi Sai, Ziqing Li, Anne C. Pivonka, Jingning Chen, and Colin M. Gray (Aug. 26, 2022). *Surveying the Landscape of Ethics-Focused Design Methods*. arXiv: 2102.08909[cs].

Chivukula, Shruthi Sai, Chris Rhys Watkins, Rhea Manocha, Jingle Chen, and Colin M. Gray (Apr. 23, 2020). „Dimensions of UX Practice that Shape Ethical Awareness“. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. New York, NY, USA: Association for Computing Machinery, pp. 1–13.

Chopra, Simran, Rachel E Clarke, Adrian K Clear, Sara Heitlinger, Ozge Dilaver, and Christina Vasiliou (Apr. 2022). „Negotiating sustainable futures in communities through participatory speculative design and experiments in living“. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New York, NY, USA: Association for Computing Machinery, pp. 1–17.

City of Barcelona (2023). *Ethical Digital Standards*. Ethical Digital Standards. URL: https://www.barcelona.cat/digitalstandards/en/init/0.1/index.html (visited on July 13, 2024).

City of London (2024). *LOTI*. London Data Ethics Service. URL: https://loti.london/resources/london-data-ethics-service/ (visited on July 13, 2024).

Cohn, Mike (2010). *Succeeding with Agile: Software Development Using Scrum*. Google-Books-ID: IdT6AgAAQBAJ. Pearson Education. 504 pp.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq (June 9, 2017). *Algorithmic decision making and the cost of fairness*. arXiv: 1701.08230[cs, stat].

Corley, M. C., R. K. Elswick, M. Gorman, and T. Clor (Jan. 2001). „Development and evaluation of a moral distress scale“. In: *Journal of Advanced Nursing* 33.2, pp. 250–256.

Corral, Luis and Ilenia Fronza (Sept. 14, 2018). „Design Thinking and Agile Practices for Software Engineering: An Opportunity for Innovation“. In: *Proceedings of the 19th Annual SIG Conference on Information Technology Education*. SIGITE ’18. New York, NY, USA: Association for Computing Machinery, pp. 26–31.

Corrêa, Nicholas Kluge, Camila Galvão, James William Santos, Carolina Del Pino, Edson Pontes Pinto, Camila Barbosa, Diogo Massmann, Rodrigo Mambrini, Luiza Galvão, and Edmund Terem (May 26, 2023). *Worldwide AI Ethics: a review of 200 guidelines and recommendations for AI governance*. arXiv: 2206.11922[cs].

Costanza-Chock, Sasha (June 3, 2018). *Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice*. Rochester, NY.

Cottrill, Caitlin D., Naomi Jacobs, Milan Markovic, and Pete Edwards (Dec. 2020). „Sensing the City: Designing for Privacy and Trust in the Internet of Things“. en. In: *Sustainable Cities and Society* 63, p. 102453.

Cowls, Josh, Thomas King, Mariarosaria Taddeo, and Luciano Floridi (May 15, 2019). *Designing AI for Social Good: Seven Essential Factors*. Rochester, NY.

Cruz, Eberth Felipe Castro da, Francisco Erivaldo Fernandes Junior, and Eduardo Drummond Sardinha (Dec. 14, 2021). „An experience in the use of SCRUM and KANBAN for project development in a Waterfall environment“. In: *Proceedings of the XX Brazilian Symposium on Software Quality*. SBQS '21. New York, NY, USA: Association for Computing Machinery, pp. 1–7.

CTH (2024). *Center for Humane Technology*. URL: `https://www.humanetech.com/who-we-are` (visited on July 2, 2024).

Culén, Alma L. and Asbjørn Følstad (Oct. 26, 2014). „Innovation in HCI: what can we learn from design thinking?“ In: *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. NordiCHI '14. New York, NY, USA: Association for Computing Machinery, pp. 849–852.

Cunningham, Jay, Gabrielle Benabdallah, Daniela Rosner, and Alex Taylor (Apr. 13, 2023). „On the Grounds of Solutionism: Ontologies of Blackness and HCI“. In: *ACM Transactions on Computer-Human Interaction* 30.2, 20:1–20:17.

Dahlgren, Kari, Sarah Pink, Yolande Strengers, Larissa Nicholls, and Jathan Sadowski (Oct. 1, 2021). „Personalization and the Smart Home: questioning techno-hedonist imaginaries“. In: *Convergence* 27.5. Publisher: SAGE Publications Ltd, pp. 1155–1169.

Davis, Jenny L., Apryl Williams, and Michael W. Yang (July 1, 2021). „Algorithmic reparation“. In: *Big Data & Society* 8.2. Publisher: SAGE Publications Ltd, p. 20539517211044808.

DDC (Oct. 5, 2021). *Toolkit: The Digital Ethics Compass*. DDC – Danish Design Center. URL: `https://ddc.dk/tools/toolkit-the-digital-ethics-compass/` (visited on Apr. 25, 2024).

Deng, Wesley Hanwen, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio (June 12, 2023). „Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice“. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. New York, NY, USA: Association for Computing Machinery, pp. 705–716.

Devon, Richard and Ibo van de Poel (2004). „Design Ethics: The Social Ethics Paradigm“. In: *International Journal of Engineering Education* 20.3, pp. 461–469.

Dignum, Virginia (Feb. 4, 2022). „Relational Artificial Intelligence“. In: *arXiv:2202.07446 [cs]*. arXiv: `2202.07446`.

Dindler, Christian, Peter Gall Krogh, Kasper Tikær, and Peter Nørregaard (Aug. 2022). „Engagements and Articulations of Ethics in Design Practice“. In: *International Journal of Dsign* 16.2, pp. 47–56.

DiSalvo, Carl (2012). *Adversarial design*. Design thinking, design theory. OCLC: ocn755213451. Cambridge, Mass: MIT Press. 145 pp.

DiSalvo, Carl (Feb. 8, 2022). *Design as Democratic Inquiry: Putting Experimental Civics into Practice*.

Dobrev, Dimiter (Oct. 3, 2012). *A Definition of Artificial Intelligence*. arXiv: `1210.1568[cs]`.

Dobrigkeit, Franziska and Danielly de Paula (Aug. 12, 2019). „Design thinking in practice: understanding manifestations of design thinking in software engineering“. In: *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ESEC/FSE 2019. New York, NY, USA: Association for Computing Machinery, pp. 1059–1069.

Dolejšová, Markéta, Cristina Ampatzidou, Lara Houston, *et al.* (June 2021). „Designing for Transformative Futures: Creative Practice, Social Change and Climate Emergency“. en. In: *Creativity and Cognition*. Virtual Event Italy: ACM, pp. 1–9.

Dörrenbächer, Judith, Matthias Laschke, Diana Löffler, Ronda Ringfort, Sabrina Großkopp, and Marc Hassenzahl (May 21, 2021). *Experiencing Utopia. A Positive Approach to Design Fiction*. arXiv: `2105.10186[cs]`.

Dot.Everyone (2024). *Consequence Scanning – an agile practice for responsible innovators – doteveryone*. URL: `https://doteveryone.org.uk/project/consequence-scanning/` (visited on Apr. 26, 2024).

Doubek, James (Nov. 16, 2017). „How Disinformation And Distortions On Social Media Affected Elections Worldwide“. In: *NPR*.

Dourish, Paul (Feb. 2004). „What we talk about when we talk about context“. en. In: *Personal and Ubiquitous Computing* 8.1, pp. 19–30.

Dourish, Paul, Janet Finlay, Phoebe Sengers, and Peter Wright (2004). „Reflective HCI: towards a critical technical practice“. In: *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI ’04*. Extended abstracts of the 2004 conference. Vienna, Austria: ACM Press, p. 1727.

Dove, Graham, Kim Halskov, Jodi Forlizzi, and John Zimmerman (May 2, 2017). „UX Design Innovation: Challenges for Working with Machine Learning as a Design Material“. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. New York, NY, USA: Association for Computing Machinery, pp. 278–288.

Drage, Eleanor, Kerry McInerney, and Jude Browne (Jan. 2, 2024). „Engineers on responsibility: feminist approaches to who’s responsible for ethical AI“. In: *Ethics and Information Technology* 26.1, p. 4.

Dunbar, W. Scott (Dec. 1, 2005). „Emotional engagement in professional ethics“. In: *Science and Engineering Ethics* 11.4, pp. 535–551.

E.S. Ruppert (2018). *Sociotechnical Imaginaries of Different Data Futures*.

EDN (2024). *Ethical Design Network*. Ethical Design Network. URL: `https://ethicaldesignnetwork.com/` (visited on July 10, 2024).

Elsden, Chris, David Chatting, Michael Duggan, Andrew Carl Dwyer, and Pip Thornton (Apr. 27, 2022). „Zoom Obscura: Counterfunctional Design for Video-Conferencing“. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI ’22. New York, NY, USA: Association for Computing Machinery, pp. 1–17.

Elsden, Chris, David Chatting, Abigail C. Durrant, Andrew Garbett, Bettina Nissen, John Vines, and David S. Kirk (May 2017). „On Speculative Enactments“. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. New York, NY, USA: Association for Computing Machinery, pp. 5386–5399.

EMAF (2024). *The Circular Design Guide*. The Circular Design Guide. URL: https://www.ellenmacarthurfoundation.org/circular-design-guide/overview (visited on July 10, 2024).

EU (Apr. 8, 2019). *Ethics guidelines for trustworthy AI*. Digital Strategy for Europe. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (visited on July 14, 2024).

EU (Apr. 21, 2021). *Proposal for a Regulation laying down harmonised rules on artificial intelligence*. Digital Strategy for Europe. URL: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (visited on July 14, 2024).

Fallman, Daniel (Apr. 1, 2010). „A different way of seeing: Albert Borgmann's philosophy of technology and human–computer interaction“. In: *AI & SOCIETY* 25.1, pp. 53–60.

Feenberg, Andrew (2009). „Democratic rationalization: Technology, power, and freedom“. In: *Readings in the philosophy of technology*. Publisher: Rowman and Littlefield Lanham, MD, pp. 139–155.

Fiesler, Casey, Shannon Morrison, and Amy S. Bruckman (May 7, 2016). „An Archive of Their Own: A Case Study of Feminist HCI and Values in Design“. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI'16: CHI Conference on Human Factors in Computing Systems. San Jose California USA: ACM, pp. 2574–2585.

Fisher, Berenice and Joan Tronto (1990). „Toward a feminist theory of caring“. In: *Circles of care: Work and identity in women's lives*. Publisher: Suny Press Albany, pp. 35–62.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar (Jan. 15, 2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*. Rochester, NY.

Floridi, Luciano, ed. (Dec. 1, 2014). *The Onlife Manifesto*. 1st ed. Springer Cham. XIV, 264.

Floridi, Luciano (Sept. 1, 2018). „Artificial Intelligence, Deepfakes and a Future of Ectypes“. In: *Philosophy & Technology* 31.3, pp. 317–321.

Floridi, Luciano (June 1, 2019). „Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical“. In: *Philosophy & Technology* 32.2, pp. 185–193.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, *et al.* (Dec. 1, 2018). „AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations“. In: *Minds and Machines* 28.4, pp. 689–707.

Forlano, Laura and Anijo Mathew (Oct. 2, 2014). „From Design Fiction to Design Friction: Speculative and Participatory Design of Values-Embedded Urban Technology“. In: *Journal of Urban Technology* 21.4, pp. 7–24.

Formosa, Paul and Malcolm Ryan (Sept. 1, 2021). „Making moral machines: why we need artificial moral agents“. In: *AI & SOCIETY* 36.3, pp. 839–851.

Foroohar, Rana (Dec. 16, 2018). *Year in a Word: Techlash*. URL: https://www.ft.com/content/76578fba-fca1-11e8-ac00-57a2a826423e (visited on June 19, 2024).

Fossa, Fabio (June 1, 2018). „Artificial moral agents: moral mentors or sensible tools?“ In: *Ethics and Information Technology* 20.2, pp. 115–126.

Frauenberger, Christopher (Feb. 2020). „Entanglement HCI The Next Wave?“ en. In: *ACM Transactions on Computer-Human Interaction* 27.1, pp. 1–27.

Frauenberger, Christopher, Marjo Rauhala, and Geraldine Fitzpatrick (Mar. 8, 2017). „In-Action Ethics“. In: *Interacting with Computers* 29.2, pp. 220–236.

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian (Sept. 23, 2016). *On the (im)possibility of fairness*. arXiv: `1609.07236[cs,stat]`.

Friedman, Batya and David Hendry (May 5, 2012). „The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations“. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’12. New York, NY, USA: Association for Computing Machinery, pp. 1145–1148.

Friedman, Batya and David G. Hendry (May 21, 2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. Google-Books-ID: 8ZiWDwAAQBAJ. MIT Press. 258 pp.

Friedman, Batya, David G. Hendry, and Alan Borning (Nov. 21, 2017). „A Survey of Value Sensitive Design Methods“. In: *Foundations and Trends® in Human–Computer Interaction* 11.2. Publisher: Now Publishers, Inc., pp. 63–125.

Friedman, Batya and Peter H. Kahn (2007). „Human Values, Ethics and Design“. In: *The human-computer interaction handbook*. 2nd ed. CRC Press, pp. 1267–1292.

Fritsch, Ester, Irina Shklovski, and Rachel Douglas-Jones (Apr. 21, 2018). „Calling for a Revolution: An Analysis of IoT Manifestos“. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. New York, NY, USA: Association for Computing Machinery, pp. 1–13.

Gardels, Nathan (Apr. 30, 2018). „Tech-lash galore“. In: *Wired*.

Garrett, Natalie, Mikhaila Friske, and Casey Fiesler (Feb. 26, 2020). „Ethics from the Start: Exploring Student Attitudes and Creating Interventions in Intro Programming Classes“. In: *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*. SIGCSE ’20. New York, NY, USA: Association for Computing Machinery, p. 1348.

Garrett, Rachael, Kristina Popova, Claudia Núñez-Pacheco, Thorhildur Asgeirsdottir, Airi Lampinen, and Kristina Höök (Apr. 19, 2023). „Felt Ethics: Cultivating Ethical Sensibility in Design Practice“. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. New York, NY, USA: Association for Computing Machinery, pp. 1–15.

Garvie, Clare and Jonathan Frankle (2016). „Facial-Recognition Software Might Have a Racial Bias Problem“. en. In: *The Atlantic*, p. 6.

Gerdes, Anne (Dec. 15, 2022). „The tech industry hijacking of the AI ethics research agenda and why we should reclaim it“. In: *Discover Artificial Intelligence* 2.1, p. 25.

Gillespie, Tarleton (July 1, 2020). „Content moderation, AI, and the question of scale“. In: *Big Data & Society* 7.2. Publisher: SAGE Publications Ltd, p. 2053951720943234.

Gillies, Marco, Rebecca Fiebrink, Atau Tanaka, *et al.* (May 7, 2016). „Human-Centred Machine Learning“. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’16. New York, NY, USA: Association for Computing Machinery, pp. 3558–3565.

Gilligan, Carol (Jan. 1, 1982). *In A Different Voice: Psychological Theory and Women's Development*. Vol. 326. 220 pp.

Girard, Monique (Aug. 1, 2011). „3. Creative Friction in a New-Media Start-Up". In: *3. Creative Friction in a New-Media Start-Up*. Princeton University Press, pp. 81–117.

Gispen, Jet (2017). *Ethics for Designers*. Ethics for Designers. URL: `https://www.ethicsfordesigners.com` (visited on Apr. 26, 2024).

Goldenfein, Jake (Jan. 29, 2019). „The Profiling Potential of Computer Vision and the Challenge of Computational Empiricism". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. New York, NY, USA: Association for Computing Machinery, pp. 110–119.

Google (2024). *Google Responsible AI Practices*. Google AI. URL: `https://ai.google/responsibility/responsible-ai-practices/` (visited on June 24, 2024).

Gould, Deborah B. (Dec. 2009). *Moving Politics: Emotion and ACT UP's Fight against AIDS*. Chicago, IL: University of Chicago Press. 536 pp.

Graf, Antonia and Marco Sonnberger (Jan. 1, 2020). „Responsibility, rationality, and acceptance: How future users of autonomous driving are constructed in stakeholders' sociotechnical imaginaries". In: *Public Understanding of Science* 29.1. Publisher: SAGE Publications Ltd, pp. 61–75.

Gray, Colin M. and Shruthi Sai Chivukula (May 2, 2019). „Ethical Mediation in UX Practice". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: Association for Computing Machinery, pp. 1–11.

Gray, Colin M., Shruthi Sai Chivukula, Thomas V Carlock, Ziqing Li, and Ja-Nae Duane (July 10, 2023). „Scaffolding Ethics-Focused Methods for Practice Resonance". In: *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. DIS '23. New York, NY, USA: Association for Computing Machinery, pp. 2375–2391.

Gray, Colin M., Ikechukwu Obi, Shruthi Sai Chivukula, *et al.* (Oct. 6, 2022). *Practitioner Trajectories of Engagement with Ethics-Focused Method Creation*. arXiv: `2210.03002[cs]`.

Green, Ben (Sept. 2021a). „Data Science as Political Action: Grounding Data Science in a Politics of Justice". In: *Journal of Social Computing* 2.3, pp. 249–265. arXiv: `1811.03435[cs]`.

Green, Ben (Sept. 2021b). „The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice". In: *Journal of Social Computing* 2.3. Conference Name: Journal of Social Computing, pp. 209–225.

Groves, Christopher (Dec. 1, 2015). „Logic of Choice or Logic of Care? Uncertainty, Technological Mediation and Responsible Innovation". In: *NanoEthics* 9.3, pp. 321–333.

Grudin, Jonathan (2009). „AI and HCI: Two Fields Divided by a Common Focus". In: *AI Magazine* 30.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1609/aimag.v30i4.2271, pp. 48–57.

H. Tan, Neilly, Brian Kinnee, Dana Langseth, Sean A. Munson, and Audrey Desjardins (Apr. 2022). „Critical-Playful Speculations with Cameras in the Home". In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New York, NY, USA: Association for Computing Machinery, pp. 1–22.

Hagendorff, Thilo (Mar. 1, 2020). „The Ethics of AI Ethics: An Evaluation of Guidelines“. In: *Minds and Machines* 30.1, pp. 99–120.

Hagendorff, Thilo (June 21, 2022). „A Virtue-Based Framework to Support Putting AI Ethics into Practice“. In: *Philosophy & Technology* 35.3, p. 55.

Hankerson, David, Andrea R. Marshall, Jennifer Booker, Houda Elmimouni, Imani Walker, and Jennifer A. Rode (May 7, 2016). „Does Technology Have Race?“ In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA '16. New York, NY, USA: Association for Computing Machinery, pp. 473–486.

Haraway, Donna (1988). „Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective“. In: *Feminist Studies* 14.3, p. 575.

Harmon, Ellie, Matthias Korn, Ann Light, and Amy Voida (June 4, 2016). „Designing Against the Status Quo“. In: *Proceedings of the 2016 ACM Conference Companion Publication on Designing Interactive Systems*. DIS '16 Companion. New York, NY, USA: Association for Computing Machinery, pp. 65–68.

Hautala, Johanna and Toni Ahlqvist (Sept. 30, 2022). „Integrating futures imaginaries, expectations and anticipatory practices: practitioners of artificial intelligence between now and future“. In: *Technology Analysis & Strategic Management* 0.0. Publisher: Routledge _eprint: https://doi.org/10.1080/09537325.2022.2130041, pp. 1–13.

Hawkins, Andrew J. (July 2022). *Tesla driver using Autopilot kills motorcyclist, prompting another NHTSA investigation*. en.

Hermann, Isabella (Feb. 1, 2023). „Artificial intelligence in fiction: between narratives and metaphors“. In: *AI & SOCIETY* 38.1, pp. 319–329.

Hoffmann, Anna Lauren (June 7, 2019). „Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse“. In: *Information, Communication & Society* 22.7. Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2019.1573912, pp. 900–915.

Holmquist, Lars Erik (June 23, 2017). „Intelligence on tap: artificial intelligence as a new design material“. In: *Interactions* 24.4, pp. 28–33.

Höök, Kristina (Nov. 13, 2018). *Designing with the Body: Somaesthetic Interaction Design*. Red. by Ken Friedman and Erik Stolterman. Design Thinking, Design Theory. Cambridge, MA, USA: MIT Press. 272 pp.

Höök, Kristina and Jonas Löwgren (Mar. 1, 2021). „Characterizing Interaction Design by Its Ideals: A Discipline in Transition“. In: *She Ji: The Journal of Design, Economics, and Innovation* 7.1, pp. 24–40.

Hort, Max, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro (June 20, 2024). „Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey“. In: *ACM J. Responsib. Comput.* 1.2, 11:1–11:52.

Husain, Syed Omer, Alex Franklin, and Dirk Roep (Mar. 2020). „The political imaginaries of blockchain projects: discerning the expressions of an emerging ecosystem“. In: *Sustainability Science* 15.2, pp. 379–394.

Hyper Island (2024). *Unintended Consequences*. HI Toolbox. URL: `https://toolbox.hyperisland.com/unintended-consequences` (visited on Apr. 26, 2024).

Ibáñez, Javier Camacho and Mónica Villas Olmeda (Dec. 1, 2022). „Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study“. In: *AI & SOCIETY* 37.4, pp. 1663–1687.

IBM (2022). *Everyday ethics for AI*. URL: https://www.ibm.com/design/ai/ethics/everyday-ethics/www.ibm.com/design/ai/ethics/everyday-ethics (visited on Apr. 25, 2024).

IDEO (July 2019). *AI & Ethics: Collaborative Activities for Designers*. URL: https://www.ideo.com/journal/ai-ethics-collaborative-activities-for-designers (visited on Apr. 25, 2024).

IDEO (2022). *How does design drive new ventures?* URL: https://www.ideo.com/question/how-does-design-drive-new-ventures (visited on Apr. 21, 2022).

IDF (2024a). *What is Circular Design*. What is Circular Design? — updated 2024. URL: https://www.interaction-design.org/literature/topics/circular-design (visited on July 10, 2024).

IDF (2024b). *What is Human-Centered Design (HCD)?* What is Human-Centered Design (HCD)? — updated 2024. URL: https://www.interaction-design.org/literature/topics/human-centered-design (visited on July 10, 2024).

IDF (2024c). *What is Universal Design?* What is Universal Design? URL: https://www.interaction-design.org/literature/topics/universal-design (visited on July 10, 2024).

Imbulana, Dilini I., Peter G. Davis, and Trisha M. Prentice (Oct. 1, 2021). „Interventions to reduce moral distress in clinicians working in intensive care: A systematic review“. In: *Intensive and Critical Care Nursing* 66, p. 103092.

Internet Society (2017). *Artificial Intelligence & Machine Learning: Policy Paper*. URL: https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/ (visited on July 14, 2024).

Irani, Lilly (2018). „"Design Thinking”: Defending Silicon Valley at the Apex of Global Labor Hierarchies | Catalyst: Feminism, Theory, Technoscience“. In: *Catalyst: Feminism, Theory, Technoscience* 4.1.

Irani, Lilly (Mar. 12, 2019). „Chasing Innovation: Making Entrepreneurial Citizens in Modern India“. In: *Chasing Innovation*. Princeton University Press.

Irani, Lilly, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter (Apr. 10, 2010). „Postcolonial computing: a lens on design and development“. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’10. New York, NY, USA: Association for Computing Machinery, pp. 1311–1320.

Irani, Lilly C. and Paul Dourish (Feb. 20, 2009). „Postcolonial interculturality“. In: *Proceedings of the 2009 international workshop on Intercultural collaboration*. IWIC ’09. New York, NY, USA: Association for Computing Machinery, pp. 249–252.

Iskander, Natasha (Sept. 5, 2018). „Design Thinking Is Fundamentally Conservative and Preserves the Status Quo“. In: *Harvard Business Review*. Section: Entrepreneurship.

Jasanoff, Sheila (2003). „In a Constitutional Moment: Science and Social Order at the Millennium“. In: *Social Studies of Science and Technology: Looking Back,* *Ahead*. Ed. by Bernward Joerges and Helga Nowotny. Sociology of the Sciences. Dordrecht: Springer Netherlands, pp. 155–180.

Jasanoff, Sheila (Sept. 2, 2015). „One. Future Imperfect: Science, Technology, and the Imaginations of Modernity“. In: *One. Future Imperfect: Science, Technology, and the Imaginations of Modernity*. University of Chicago Press, pp. 1–33.

Jasanoff, Sheila and Sang-Hyun Kim (June 2009). „Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea“. In: *Minerva* 47.2, pp. 119–146.

Jobin, Anna, Marcello Ienca, and Effy Vayena (Sept. 2019). „The global landscape of AI ethics guidelines“. In: *Nature Machine Intelligence* 1.9. Number: 9 Publisher: Nature Publishing Group, pp. 389–399.

Johnson, Deborah G. (Dec. 15, 2000). *Computer Ethics (3rd Edition)*. 3nd edition. Open Library ID: OL9288025M. Prentice Hall. 256 pp.

Johnson, Khari (2022). „How Wrongful Arrests Based on AI Derailed 3 Men's Lives“. en-US. In: *Wired* (). Section: tags.

Jones, Mirabelle, Christina Neumayer, and Irina Shklovski (Apr. 19, 2023). „Embodying the Algorithm: Exploring Relationships with Large Language Models Through Artistic Performance“. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. New York, NY, USA: Association for Computing Machinery, pp. 1–24.

Jordan, Jennifer (Jan. 1, 2009). „A Social Cognition Framework for Examining Moral Awareness in Managers and Academics“. In: *Journal of Business Ethics* 84.2, pp. 237–258.

Joyce, Kelly, Laurel Smith-Doerr, Sharla Alegria, Susan Bell, Taylor Cruz, Steve G. Hoffman, Safiya Umoja Noble, and Benjamin Shestakofsky (Jan. 1, 2021). „Toward a Sociology of Artificial Intelligence: A Call for Research on Inequalities and Structural Change“. In: *Socius* 7. Publisher: SAGE Publications, p. 2378023121999581.

Kahn, Batya Friedman {and} Peter H. and Jr (2007). „HUMAN VALUES, ETHICS, AND DESIGN“. In: *The Human-Computer Interaction Handbook*. 2nd ed. Num Pages: 26. CRC Press.

Kapania, Shivani, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan (Apr. 27, 2022). „"Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India“. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New York, NY, USA: Association for Computing Machinery, pp. 1–18.

Kaplan, Andreas and Michael Haenlein (Jan. 1, 2019). „Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence“. In: *Business Horizons* 62.1, pp. 15–25.

Kawall, Jason (2008). „In Defense of the Primary of the Virtues“. In: *Journal of Ethics & Social Philosophy* 3.2, [i]–21.

Key, Cayla, Fiona Browne, Nick Taylor, and Jon Rogers (May 2021). „Proceed with Care: Reimagining Home IoT Through a Care Perspective“. en. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama Japan: ACM, pp. 1–15.

Khan, Arif Ali, Sher Badshah, Peng Liang, Muhammad Waseem, Bilal Khan, Aakash Ahmad, Mahdi Fahmideh, Mahmood Niazi, and Muhammad Azeem Akbar (June 13, 2022). „Ethics of AI: A Systematic Literature Review of Principles and Challenges“. In: *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022*. EASE ’22. New York, NY, USA: Association for Computing Machinery, pp. 383–392.

Kordzadeh, Nima and Maryam Ghasemaghaei (May 2022). „Algorithmic bias: review, synthesis, and future research directions“. EN. In: *European Journal of Information Systems*. Publisher: Taylor & Francis.

Kötteritzsch, Anna, Michael Koch, and Susanne Wallrafen (Sept. 2016). „Expand your comfort zone! smart urban objects to promote safety in public spaces for older adults“. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. UbiComp ’16. New York, NY, USA: Association for Computing Machinery, pp. 1399–1407.

Kow, Yong Ming and Caitlin Lustig (Apr. 1, 2018). „Imaginaries and Crystallization Processes in Bitcoin Infrastructuring“. In: *Computer Supported Cooperative Work (CSCW)* 27.2, pp. 209–232.

Kranakis, Eda (2004). „Fixing the Blame: Organizational Culture and the Quebec Bridge Collapse“. In: *Technology and Culture* 45.3. Publisher: [The Johns Hopkins University Press, Society for the History of Technology], pp. 487–518.

Kronqvist, Aila and Rebekah Ann Rousi (Jan. 2, 2023). „A quick review of ethics, design thinking, gender, and AI development“. In: *International Journal of Design Creativity and Innovation* 11.1. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21650349.2022.2136 pp. 62–79.

Krumm, John (2007). „Inference Attacks on Location Tracks“. en. In: *Pervasive Computing*. Ed. by Anthony LaMarca, Marc Langheinrich, and Khai N. Truong. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 127–143.

Lampinen, Airi, Moira McGregor, Rob Comber, and Barry Brown (Nov. 1, 2018). „Member-Owned Alternatives: Exploring Participatory Forms of Organising with Cooperatives“. In: *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW), 100:1–100:19.

Larrabee, Mary Jeanne (1993). „Gender and Moral Development: A Challenge for Feminist Theory“. In: *An Ethic of Care*. Num Pages: 14. Routledge.

Latour, Bruno (1992). „Where are the missing masses? The sociology of a few mundane artifacts“. In: *Shaping technology/building society: Studies in sociotechnical change* 1. Publisher: Mit Press Cambridge, MA, pp. 225–258.

Lauer, Dave (Feb. 1, 2021). „You cannot have AI ethics without ethics“. In: *AI and Ethics* 1.1, pp. 21–25.

Law, John, Evelyn Ruppert, and Mike Savage (2013). „The double social life of methods“. In: *CRESC working paper series*. CRESC working paper series Working Paper No. 95 (Milton Keynes, UK: Centre for Research on Socio-Cultural Change, The Open University).

Le Dantec, Christopher A., Erika Shehan Poole, and Susan P. Wyche (Apr. 4, 2009). „Values as lived experience: evolving value sensitive design in support of value discovery“. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing*

Systems. CHI ’09. New York, NY, USA: Association for Computing Machinery, pp. 1141–1150.

Lee, Minsoo, Yoonsik Uhm, Zion Hwang, Yong Kim, Joohyung Jo, and Sehyun Park (Oct. 2007). „A Ubiquitous Computing Network Framework for Assisting People in Urban Areas“. In: *32nd IEEE Conference on Local Computer Networks (LCN 2007)*. ISSN: 0742-1303, pp. 215–216.

Lewis, Paul (Oct. 6, 2017). „’Our minds can be hijacked’: the tech insiders who fear a smartphone dystopia“. In: *The Guardian*.

Liao, Q. Vera, Daniel Gruen, and Sarah Miller (Apr. 21, 2020). „Questioning the AI: Informing Design Practices for Explainable AI User Experiences“. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. New York, NY, USA: Association for Computing Machinery, pp. 1–15.

Light, Ann (Sept. 1, 2011). „HCI as heterodoxy: Technologies of identity and the queering of interaction with computers“. In: *Interacting with Computers* 23.5, pp. 430–438.

Light, Ann (Dec. 2021). „Collaborative speculation: Anticipation, inclusion and designing counterfactual futures for appropriation“. en. In: *Futures* 134, p. 102855.

Light, Ann and Yoko Akama (2014). „Structuring future social relations: the politics of care in participatory practice“. en. In: *Proceedings of the 13th Participatory Design Conference on Research Papers - PDC ’14*. Windhoek, Namibia: ACM Press, pp. 151–160.

Lindberg, Sharon, Petter Karlström, and Sirkku Männikkö Barbutiu (June 12, 2023). „Cultivating ethics with professional designers“. In: *Nordes Conference Series*.

Lindberg, Sharon, Chiara Rossitto, Ola Knutsson, Petter Karlström, and Sirkku Männikkö Barbutiu (Feb. 21, 2024). „Doing Good Business? Design Leaders’ Perspectives on Ethics in Design“. In: *Proceedings of the ACM on Human-Computer Interaction* 8 (GROUP), 2:1–2:22.

Lindtner, Silvia, Ken Anderson, and Paul Dourish (Feb. 11, 2012). „Cultural appropriation: information technologies as sites of transnational imagination“. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. CSCW ’12. New York, NY, USA: Association for Computing Machinery, pp. 77–86.

Lockton, Dan, Michelle Chou, Aadya Krishnaprasad, Deepika Dixit, Stefania La Vattiata, Jisoo Shon, Matt Geiger, and Tammar Zea-Wolfson (Dec. 10, 2019). *Metaphors and imaginaries in design research for change*. DR4C: Design Research for Change Symposium. Design Museum, London.

Lopes, João, Marcia Gusmao, Rodrigo Souza, Patricia Davet, Alexandre Souza, Cristiano Costa, Jorge Barbosa, Ana Pernas, Adenauer Yamin, and Claudio Geyer (Nov. 2013). „Towards a distributed architecture for context-aware mobile applications in UbiComp“. In: *Proceedings of the 19th Brazilian symposium on Multimedia and the web*. WebMedia ’13. New York, NY, USA: Association for Computing Machinery, pp. 43–50.

Löwgren, Jonas and Erik Stolterman (2007). *Thoughtful Interaction Design: A Design Perspective on Information Technology*. Google-Books-ID: Qr74DwAAQBAJ. MIT Press. 213 pp.

Luján Escalante, Maria, Luke Moffat, and Monika Büscher (June 25, 2022). „Ethics through design“. In: *DRS Biennial Conference Series*.

Lustig, Caitlin (Nov. 7, 2019). „Intersecting Imaginaries: Visions of Decentralized Autonomous Systems“. In: *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW), pp. 1–27.

Madaio, Michael A., Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach (Apr. 23, 2020). „Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI“. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. New York, NY, USA: Association for Computing Machinery, pp. 1–14.

Mainsah, Henry and Andrew Morrison (Oct. 6, 2014). „Participatory design through a cultural lens: insights from postcolonial theory“. In: *Proceedings of the 13th Participatory Design Conference: Short Papers, Industry Cases, Workshop Descriptions, Doctoral Consortium papers, and Keynote abstracts - Volume 2*. PDC ’14. New York, NY, USA: Association for Computing Machinery, pp. 83–86.

Manders-Huits, Noëmi (June 1, 2011). „What Values in Design? The Challenge of Incorporating Moral Values into Design“. In: *Science and Engineering Ethics* 17.2, pp. 271–287.

Mark, Gloria, Shamsi Iqbal, Mary Czerwinski, and Paul Johns (Feb. 2015). „Focused, Aroused, but so Distractible: Temporal Perspectives on Multitasking and Communications“. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW ’15. New York, NY, USA: Association for Computing Machinery, pp. 903–916.

Markham, Annette (Feb. 1, 2021). „The limits of the imaginary: Challenges to intervening in future speculations of memory, data, and algorithms“. In: *New Media & Society* 23.2. Publisher: SAGE Publications, pp. 382–405.

Martin, Aryn, Natasha Myers, and Ana Viseu (Oct. 1, 2015). „The politics of care in technoscience“. In: *Social Studies of Science* 45.5, pp. 625–641.

Maze, Ramia (2019). „Politics of Designing Visions of the Future“. In: Accepted: 2019-08-15T08:25:36Z Publisher: Tamkang University.

McCarthy, John and Peter Wright (2004). *Technology as Experience*. English. Cambridge, Mass: The MIT Press.

McCarthy, John and Peter Wright (Nov. 1, 2005). „Putting ‘felt-life’ at the centre of human–computer interaction (HCI)“. In: *Cognition, Technology & Work* 7.4, pp. 262–271.

McEvers, Kelly (July 10, 2017). *Tech Design Ethicist Works To Raise Awareness Of Internet Addiction*. In collab. with Tristan Harris.

McNamara, Andrew, Justin Smith, and Emerson Murphy-Hill (Oct. 26, 2018). „Does ACM’s code of ethics change ethical decision making in software development?“ In: *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ESEC/FSE 2018. New York, NY, USA: Association for Computing Machinery, pp. 729–733.

Medler, Ben and Brian Magerko (2010). „The implications of improvisational acting and role-playing on design methodologies“. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 483–492.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan (July 13, 2021). „A Survey on Bias and Fairness in Machine Learning“. In: *ACM Comput. Surv.* 54.6, 115:1–115:35.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan (Jan. 25, 2022). *A Survey on Bias and Fairness in Machine Learning*. arXiv: 1908.09635[cs].

Metcalf, Jacob, Emanuel Moss, and danah boyd danah (2019). „Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics“. In: *Social Research: An International Quarterly* 86.2. Publisher: Johns Hopkins University Press, pp. 449–476.

Microsoft (Feb. 27, 2023). *What is Microsoft's Approach to AI? | Microsoft Source*. Microsoft News. URL: https://news.microsoft.com/source/features/ai/microsoft-approach-to-ai/ (visited on Sept. 8, 2023).

Microsoft (2024). *Microsoft Inclusive Design*. URL: https://inclusive.microsoft.design/ (visited on Apr. 26, 2024).

Miller, Jessica K., Batya Friedman, Gavin Jancke, and Brian Gill (Nov. 2007). „Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system“. In: *Proceedings of the 2007 ACM International Conference on Supporting Group Work*. GROUP '07. New York, NY, USA: Association for Computing Machinery, pp. 281–290.

Miller, Thaddeus R. (July 2, 2020). „Imaginaries of Sustainability: The Techno-Politics of Smart Cities“. In: *Science as Culture* 29.3. Publisher: Routledge _eprint: https://doi.org/10.1080/09505431.2019.1705273, pp. 365–387.

Mittelstadt, Brent (Nov. 2019). „Principles alone cannot guarantee ethical AI“. In: *Nature Machine Intelligence* 1.11. Number: 11 Publisher: Nature Publishing Group, pp. 501–507.

Mittelstadt, Brent, Sandra Wachter, and Chris Russell (Jan. 20, 2023). *The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default*. Rochester, NY.

Mol, Annemarie (2008). *The Logic of Care: Health and the Problem of Patient Choice*. 1 edition. London ; New York: Routledge. 160 pp.

Mol, Annemarie, Ingunn Moser, and Jeannette Pols, eds. (Dec. 31, 2010). *Care in Practice: On Tinkering in Clinics, Homes and Farms*. transcript Verlag.

Möllering, Guido (2006). *Trust: reason, routine, reflexivity*. en. Amsterdam: Elsevier.

Monteiro, Mike (2019). *Ruined by design: How designers destroyed the world, and what we can do to fix it*. Mule Design.

Moor, James H. (1985). „What Is Computer Ethics?*“. In: *Metaphilosophy* 16.4. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9973.1985.tb00173.x, pp. 266–275.

Moor, James H. (Mar. 1, 1999). „Just consequentialism and computing“. In: *Ethics and Information Technology* 1.1, pp. 61–65.

Morley, Georgina, Jonathan Ives, Caroline Bradbury-Jones, and Fiona Irvine (May 2019). „What is 'moral distress'? A narrative synthesis of the literature“. In: *Nursing Ethics* 26.3, pp. 646–662.

Morley, Jessica, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander, and Luciano Floridi (June 1, 2021). „Ethics as a Service: A Pragmatic Operationalisation of AI Ethics“. In: *Minds and Machines* 31.2, pp. 239–256.

Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal (Aug. 1, 2020). „From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices“. In: *Science and Engineering Ethics* 26.4. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 4 Publisher: Springer Netherlands, pp. 2141–2168.

Morley, Jessica, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi (Feb. 1, 2023). „Operationalising AI ethics: barriers, enablers and next steps“. In: *AI & SOCIETY* 38.1, pp. 411–423.

Munn, Luke (Aug. 23, 2022). „The uselessness of AI ethics“. In: *AI and Ethics*.

Musa Giuliano, Roberto (Dec. 1, 2020). „Echoes of myth and magic in the language of Artificial Intelligence“. In: *AI & SOCIETY* 35.4, pp. 1009–1024.

Nägele, Larissa Vivian, Merja Ryöppy, and Danielle Wilde (Sept. 2018). „PDFi: participatory design fiction with vulnerable users“. In: *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*. NordiCHI ’18. New York, NY, USA: Association for Computing Machinery, pp. 819–831.

Nardi, Bonnie and Yong Ming Kow (June 5, 2010). „Digital imaginaries: How we know what we (think we) know about Chinese gold farming“. In: *First Monday*.

Nelson, Harold G. and Erik Stolterman (Aug. 29, 2014). *The Design Way, second edition: Intentional Change in an Unpredictable World*. Google-Books-ID: lr34DwAAQBAJ. MIT Press. 297 pp.

Nissenbaum, Helen (2004). „Privacy as Contextual Integrity Symposium - Technology, Values, and the Justice System“. In: *Washington Law Review* 79.1, pp. 119–158.

Noddings, Nel (Sept. 2013). *Caring: A Relational Approach to Ethics and Moral Education, Updated*. 2nd ed. 256 pp.

Novelli, Claudio, Mariarosaria Taddeo, and Luciano Floridi (Feb. 7, 2023). „Accountability in artificial intelligence: what it is and how it works“. In: *AI & SOCIETY*.

O’Neil, Cathy (Sept. 6, 2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Google-Books-ID: NgEwCwAAQBAJ. Crown. 290 pp.

OECD (2024). *OECD Legal Instruments*. Recommendation of the Council on Artificial Intelligence. URL: `https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449` (visited on July 14, 2024).

Oomen, Jeroen, Jesse Hoffman, and Maarten A. Hajer (May 1, 2022). „Techniques of futuring: On how imagined futures become socially performative“. In: *European Journal of Social Theory* 25.2. Publisher: SAGE Publications Ltd, pp. 252–270.

OPEN (June 28, 2021). *The Data Ethics Canvas*. The ODI. URL: `https://theodi.org/insights/tools/the-data-ethics-canvas-2021/` (visited on Apr. 26, 2024).

Ortega-Bolaños, Ricardo, Joshua Bernal-Salcedo, Mariana Germán Ortiz, Julian Galeano Sarmiento, Gonzalo A. Ruz, and Reinel Tabares-Soto (Apr. 5, 2024). „Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems“. In: *Artificial Intelligence Review* 57.5, p. 110.

Oudshoorn, Nelly, Els Rommes, and Marcelle Stienstra (Jan. 1, 2004). „Configuring the User as Everybody: Gender and Design Cultures in Information and Communication Technologies“. In: *Science, Technology, & Human Values* 29.1. Publisher: SAGE Publications Inc, pp. 30–63.

Ozkaramanli, Deger, Merlijn Smits, Maaike Harbers, Gabriele Ferri, Michael Nagenborg, and Ibo van de Poel (June 23, 2024). „Navigating ethics-informed methods at the intersection of design and philosophy of technology“. In: *DRS Biennial Conference Series*.

Pair (2024). *People + AI Research*. URL: `https://pair.withgoogle.com` (visited on June 24, 2024).

Paltieli, Guy (Dec. 1, 2022). „The political imaginary of National AI Strategies“. In: *AI & SOCIETY* 37.4, pp. 1613–1624.

Parvin, Nassim and Anne Pollock (Aug. 1, 2020). „Unintended by Design: On the Political Uses of “Unintended Consequences”“. In: *Engaging Science, Technology, and Society* 6, p. 320.

Pereira, David (July 6, 2020). *Bias in AI: Much more than a Data problem*. Medium. URL: `https://towardsdatascience.com/bias-in-ai-much-more-than-a-data-problem-de6ef950c848` (visited on May 21, 2021).

Phan, Thao, Jake Goldenfein, Monique Mann, and Declan Kuch (Nov. 4, 2021). „Economies of Virtue: The Circulation of ‘Ethics’ in Big Tech“. In: *Science as Culture*. Publisher: Routledge.

Pillai, Ajit G., Thida Sachathep, and Naseem Ahmadpour (Oct. 8, 2022). „Exploring the experience of ethical tensions and the role of community in UX practice“. In: *Nordic Human-Computer Interaction Conference*. NordiCHI ’22. New York, NY, USA: Association for Computing Machinery, pp. 1–13.

Pink, Sarah (Nov. 26, 2020). „Anthropology in an uncertain world“. In: *Why the World Needs Anthropologists*. Pages: 56-70 Publication Title: Why the World Needs Anthropologists. Routledge, pp. 56–70.

Pink, Sarah (May 28, 2021). „The Ethnographic Hunch“. In.

Pink, Sarah (2022). „Trust, Ethics and Automation: Anticipatory imaginaries in everyday life“. In: *Everyday Automation: Experiencing and Anticipating Emerging Technologies*. Routledge, pp. 44–58.

Poel, Ibo van de (June 1, 2016). „An Ethical Framework for Evaluating Experimental Technology“. In: *Science and Engineering Ethics* 22.3, pp. 667–686.

Poel, Ibo van de and A. C. van Gorp (May 1, 2006). „The Need for Ethical Reflection in Engineering Design: The Relevance of Type of Design and Design Hierarchy“. In: *Science, Technology, & Human Values* 31.3. Publisher: SAGE Publications Inc, pp. 333–360.

Popova, Kristina, Claudia Figueras, Kristina Höök, and Airi Lampinen (2023). „Who Should Act? Distancing and Vulnerability in Technology Practitioners’ Accounts of Ethical Responsibility“. In.

Popova, Kristina, Rachael Garrett, Claudia Núñez-Pacheco, Airi Lampinen, and Kristina Höök (Apr. 29, 2022). „Vulnerability as an ethical stance in soma design processes“. In: *Proceedings of the 2022 CHI Conference on Human Factors* *in Computing Systems*. CHI ’22. New York, NY, USA: Association for Computing Machinery, pp. 1–13.

Powell, Alison B (Dec. 1, 2018). „Moral Orders in Contribution Cultures“. In: *Communication, Culture and Critique* 11.4, pp. 513–529.

Powell, Alison B, Funda Ustek-Spilda, Sebastián Lehuedé, and Irina Shklovski (July 1, 2022). „Addressing ethical gaps in ‘Technology for Good’: Foregrounding care and capabilities“. In: *Big Data & Society* 9.2. Publisher: SAGE Publications Ltd, p. 20539517221113774.

Prado de O. Martins, Luiza and Pedro Oliveira (Aug. 2017). *Questioning the “critical” in Speculative & Critical Design*. en. Medium.

Pschetz, Larissa, Kruakae Pothong, and Chris Speed (May 2019). „Autonomous Distributed Energy Systems: Problematising the Invisible through Design, Drama and Deliberation“. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. New York, NY, USA: Association for Computing Machinery, pp. 1–14.

PWC (Mar. 7, 2024). *Responsible AI Toolkit*. PwC. URL: `https://www.pwc.com/sg/en/services/reimagine-digital/data-optimisation/what-is-responsible-ai.html` (visited on Apr. 25, 2024).

Raji, Inioluwa Deborah, Morgan Klaus Scheuerman, and Razvan Amironesei (Mar. 1, 2021). „You Can’t Sit With Us: Exclusionary Pedagogy in AI Ethics Education“. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. New York, NY, USA: Association for Computing Machinery, pp. 515–525.

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes (Jan. 27, 2020). „Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing“. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, pp. 33–44.

Rattay, Sonja, Robert Collins, Aditi Surana, *et al.* (July 10, 2023). „Sensing Care Through Design: A Speculative Role-play Approach to "Living with" Sensor-supported Care Networks“. In: *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. DIS ’23. New York, NY, USA: Association for Computing Machinery, pp. 1660–1675.

Reeves, Stuart, Murray Goulden, and Robert Dingwall (July 1, 2016). „The Future as a Design Problem“. In: *Design Issues* 32.3, pp. 6–17.

Reijers, Wessel, David Wright, Philip Brey, Karsten Weber, Rowena Rodrigues, Declan O’Sullivan, and Bert Gordijn (Oct. 1, 2018). „Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendations“. In: *Science and Engineering Ethics* 24.5, pp. 1437–1481.

Reynolds, Scott J and Jared A Miller (Dec. 1, 2015). „The recognition of moral issues: moral awareness, moral sensitivity and moral attentiveness“. In: *Current Opinion in Psychology*. Morality and ethics 6, pp. 114–117.

Reynolds, Scott J., Bradley P. Owens, and Alex L. Rubenstein (Apr. 1, 2012). „Moral Stress: Considering the Nature and Effects of Managerial Moral Uncertainty“. In: *Journal of Business Ethics* 106.4, pp. 491–502.

Richardson, Rashida (May 20, 2021). *Racial Segregation and the Data-Driven Society: How Our Failure to Reckon with Root Causes Perpetuates Separate and Unequal Realities*. SSRN Scholarly Paper ID 3850317. Rochester, NY: Social Science Research Network.

Robbins, Holly, Elisa Giaccardi, and Elvin Karana (Oct. 23, 2016). „Traces as an Approach to Design for Focal Things and Practices“. In: *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. NordiCHI ’16. New York, NY, USA: Association for Computing Machinery, pp. 1–10.

Roberge, Jonathan, Marius Senneville, and Kevin Morin (Jan. 2020). „How to translate artificial intelligence? Myths and justifications in public discourse“. In: *Big Data & Society* 7.1, p. 205395172091996.

Rosenberg, Matthew (Mar. 19, 2018). „Cambridge Analytica, Trump-Tied Political Firm, Offered to Entrap Politicians“. In: *The New York Times*.

Rotondo, Amanda and Nathan G. Freier (Apr. 10, 2010). „The problem of defining values for design: a lack of common ground between industry and academia?“ In: *CHI ’10 Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’10. New York, NY, USA: Association for Computing Machinery, pp. 4183–4188.

Ruckenstein, Minna and Linda Lisa Maria Turunen (Sept. 19, 2019). „Re-humanizing the platform: Content moderators and the logic of care“. In: *New Media & Society*, p. 1461444819875990.

Sadek, Malak, Emma Kallina, Thomas Bohné, Céline Mougenot, Rafael A. Calvo, and Stephen Cave (Feb. 19, 2024). „Challenges of responsible AI in practice: scoping review and recommended actions“. In: *AI & SOCIETY*.

Sadowski, Jathan and Roy Bendor (May 1, 2019). „Selling Smartness: Corporate Narratives and the Smart City as a Sociotechnical Imaginary“. In: *Science, Technology, & Human Values* 44.3. Publisher: SAGE Publications Inc, pp. 540–563.

Sartori, Laura and Giulia Bocca (Mar. 26, 2022). „Minding the gap(s): public perceptions of AI and socio-technical imaginaries“. In: *AI & SOCIETY*.

Schechner, Sam (Feb. 22, 2019). „You Give Apps Sensitive Personal Information. Then They Tell Facebook.“ In: *Wall Street Journal*.

Schlesinger, Ari, W. Keith Edwards, and Rebecca E. Grinter (2021). *Intersectional HCI | Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. URL: `https://dl.acm.org/doi/10.1145/3025453.3025766` (visited on Sept. 14, 2021).

Schwennesen, Nete (2019). „Algorithmic assemblages of care: imaginaries, epistemologies and repair work“. In: *Sociology of Health & Illness* 41 (S1). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9566.12900, pp. 176–192.

Seberger, John S., Marissel Llavore, Nicholas Nye Wyant, Irina Shklovski, and Sameer Patil (May 7, 2021). „Empowering Resignation: There’s an App for That“. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. New York, NY, USA: Association for Computing Machinery, pp. 1–18.

Shahriari, Kyarash and Mana Shahriari (2017). *IEEE standard review — Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems*.

Shilton, Katie (May 1, 2013). „Values Levers: Building Ethics into Design“. In: *Science, Technology, & Human Values* 38.3. Publisher: SAGE Publications Inc, pp. 374–397.

Shilton, Katie (Mar. 1, 2018a). „Engaging Values Despite Neutrality: Challenges and Approaches to Values Reflection during the Design of Internet Infrastructure“. In: *Science, Technology, & Human Values* 43.2. Publisher: SAGE Publications Inc, pp. 247–269.

Shilton, Katie (July 15, 2018b). „Values and Ethics in Human-Computer Interaction“. In: *Foundations and Trends® in Human–Computer Interaction* 12.2. Publisher: Now Publishers, Inc., pp. 107–171.

Shklovski, Irina (Jan. 2021). *Addressing Ethical Dilemmas in AI: Listening to Engineers Report*.

Shklovski, Irina and Carolina Némethy (Jan. 2, 2023). „Nodes of certainty and spaces for doubt in AI ethics for engineers“. In: *Information, Communication & Society* 26.1. Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2021.2014547, pp. 37–53.

Sicart, Miguel and Irina Shklovski (July 3, 2020). „’Pataphysical Software: (Ridiculous) Technological Solutions for Imaginary Problems“. In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. DIS ’20. New York, NY, USA: Association for Computing Machinery, pp. 1859–1871.

Simon, Judith, Pak-Hang Wong, and Gernot Rieder (Dec. 18, 2020). „Algorithmic bias and the Value Sensitive Design approach“. In: *Internet Policy Review* 9.4.

Skirpan, Michael, Maggie Oates, Daragh Byrne, Robert Cunningham, and Lorrie Faith Cranor (Feb. 23, 2022). „Is a privacy crisis experienced, a privacy crisis avoided?“ In: *Communications of the ACM* 65.3, pp. 26–29.

Skjølsvold, Tomas Moe and Carmel Lindkvist (Sept. 1, 2015). „Ambivalence, designing users and user imaginaries in the European smart grid: Insights from an interdisciplinary demonstration project“. In: *Energy Research & Social Science*. Special Issue on Smart Grids and the Social Sciences 9, pp. 43–50.

Smith, Scott and Madeline Ashby (Sept. 10, 2020). *How to Future: Leading and Sense-making in an Age of Hyperchange*. Google-Books-ID: 2Sv7DwAAQBAJ. Kogan Page Publishers. 249 pp.

Speed, Chris, Bettina Nissen, Larissa Pschetz, Dave Murray-Rust, Hadi Mehrpouya, and Shaune Oosthuizen (Apr. 1, 2019). „Designing New Socio-Economic Imaginaries“. In: *The Design Journal* 22 (sup1), pp. 2257–2261.

Sreenivasan, Gopal (Jan. 1, 2002). „Errors about Errors: Virtue Theory and Trait Attribution“. In: *Mind* 111.441, pp. 47–68.

Srinivasan, Ramesh (2017). *Whose Global Village?: Rethinking How Technology Shapes Our World*. NYU Press.

Stark, Luke (Dec. 2021). „Apologos: A Lightweight Design Method for Sociotechnical Inquiry“. In: *Journal of Social Computing* 2.4, pp. 297–308.

Steen, Marc (May 1, 2015). „Upon Opening the Black Box and Finding It Full: Exploring the Ethics in Design Practices“. In: *Science, Technology, & Human Values* 40.3. Publisher: SAGE Publications Inc, pp. 389–420.

Stewart, Alex (1998). *The Ethnographers Method*. SAGE Publications, Inc.

Stilgoe, Jack, Richard Owen, and Phil Macnaghten (Nov. 1, 2013). „Developing a framework for responsible innovation“. In: *Research Policy* 42.9, pp. 1568–1580.

Stone, Peter, Rodney Brooks, Erik Brynjolfsson, *et al.* (2022). „Artificial Intelligence and Life in 2030: The One Hundred Year Study on Artificial Intelligence“. In: Publisher: arXiv Version Number: 1.

Strong, David and Eric Higgs (Dec. 15, 2010). „1. Borgmann’s Philosophy of Technology“. In: *1. Borgmann’s Philosophy of Technology*. University of Chicago Press, pp. 17–37.

Su, Norman Makoto, Amanda Lazar, and Lilly Irani (Apr. 22, 2021). „Critical Affects: Tech Work Emotions Amidst the Techlash“. In: *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1), 179:1–179:27.

Suchman, Lucy A. (Nov. 26, 1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Google-Books-ID: AJ_eBJtHxmsC. Cambridge University Press. 224 pp.

Sutherland, Jeff and Ken Schwaber (2024). *Scrum Guides*. scrumguides.org. URL: `https://scrumguides.org/` (visited on June 24, 2024).

Tada.city (2023). Tada.city. URL: `https://tada.city/` (visited on Mar. 16, 2023).

Tangible (2024). *Ethical Compass*. Tangible. URL: `https://tangible.is/en/ethical-compass` (visited on Apr. 26, 2024).

, tanya snook tanya (Oct. 18, 2021). *UX design has a dirty secret*. Fast Company. URL: `https://www.fastcompany.com/90686473/ux-design-has-a-dirty-secret` (visited on Feb. 2, 2022).

Taplin, Jonathan (Mar. 2017). *Move Fast and Break Things: How Facebook, Google, and Amazon Cornered Culture and Undermined Democracy*. USA: Atlantic/Little, Brown.

Taylor, Alex S. (May 7, 2011). „Out there“. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’11. New York, NY, USA: Association for Computing Machinery, pp. 685–694.

Taylor, Charles (2002). „Modern Social Imaginaries“. In: *Public Culture* 14.1. Publisher: Duke University Press, pp. 91–124.

Taylor, Linnet and Lina Dencik (Apr. 10, 2020). „Constructing Commercial Data Ethics“. In: *Technology and Regulation* 2020, pp. 1–10.

Thoughtworks (2021). *Responsible tech playbook*.

Thylstrup, Nanna Bonde (July 1, 2019). „Data out of place: Toxic traces and the politics of recycling“. In: *Big Data & Society* 6.2, p. 2053951719875479.

Tolmeijer, Suzanne, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein (Dec. 30, 2021). „Implementations in Machine Ethics: A Survey“. In: *ACM Comput. Surv.* 53.6, 132:1–132:38.

Tran O’Leary, Jasper, Sara Zewde, Jennifer Mankoff, and Daniela K. Rosner (May 2, 2019). „Who Gets to Future? Race, Representation, and Design Methods in Africatown“. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing*

Systems. CHI '19. New York, NY, USA: Association for Computing Machinery, pp. 1–13.

Tronto, Joan C. (July 2010). „Creating Caring Institutions: Politics, Plurality, and Purpose". In: *Ethics and Social Welfare* 4.2, pp. 158–171.

Turner, Fred (2010). *From counterculture to cyberculture: Stewart Brand, the Whole Earth Network, and the rise of digital utopianism*. University of Chicago Press.

Umbrello, Steven and Ibo van de Poel (Aug. 1, 2021). „Mapping value sensitive design onto AI for social good principles". In: *AI and Ethics* 1.3, pp. 283–296.

Ustek-Spilda, Funda, Alison Powell, and Selena Nemorin (July 1, 2019). „Engaging with ethics in Internet of Things: Imaginaries in the social milieu of technology developers". In: *Big Data & Society* 6.2. Publisher: SAGE Publications Ltd, p. 2053951719879468.

Vakil, Sepehr (Mar. 1, 2018). „Ethics, Identity, and Political Vision: Toward a Justice-Centered Approach to Equity in Computer Science Education". In: *Harvard Educational Review* 88.1, pp. 26–52.

Vakkuri, Ville, Kai-Kristian Kemell, Joni Kultanen, and Pekka Abrahamsson (July 2020). „The Current State of Industrial Practice in Artificial Intelligence Ethics". In: *IEEE Software* 37.4. Conference Name: IEEE Software, pp. 50–57.

Vallès-Peris, Núria and Miquel Domènech (Sept. 1, 2020). „Roboticists' Imaginaries of Robots for Care: The Radical Imaginary as a Tool for an Ethical Discussion". In: *Engineering Studies* 12.3. Publisher: Routledge _eprint: https://doi.org/10.1080/19378629.2020.1821695, pp. 157–176.

Vallor, Shannon (Sept. 22, 2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.

Vallor, Shannon (June 22, 2018). *An Ethical Toolkit for Engineering/Design Practice*. Markkula Center for Applied Ethics. URL: `https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/` (visited on July 14, 2024).

Vanhée, Loïs and Melania Borit (May 5, 2022). „Viewpoint: Ethical By Designer - How to Grow Ethical Designers of Artificial Intelligence". In: *Journal of Artificial Intelligence Research* 73.

Verbeek, Peter-Paul (2002). „Devices of Engagement: On Borgmann's Philosophy of Information and Technology". In: *Techné: Research in Philosophy and Technology* 6.1. Publisher: Society for Philosophy and Technology, pp. 48–63.

Vesti, Helle, Linda N. Laursen, and Christian Tollestrup (June 23, 2024). „Past, present and future of design ethics". In: *DRS Biennial Conference Series*.

Vines, John, Tess Denman-Cleaver, Paul Dunphy, Peter Wright, and Patrick Olivier (Apr. 2014). „Experience design theatre: exploring the role of live theatre in scaffolding design dialogues". en. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Toronto Ontario Canada: ACM, pp. 683–692.

Wachter-Boettcher, Sara (2017). *Technically wrong: sexist apps, biased algorithms, and other threats of toxic tech*. WW Norton & Company.

Wadley, Greg, Vassilis Kostakos, Peter Koval, Wally Smith, Sarah Webber, Anna Cox, James J. Gross, Kristina Höök, Regan Mandryk, and Petr Slovak (Apr. 28, 2022). „The Future of Emotion in Human-Computer Interaction". In: *Extended Abstracts* *of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI EA ’22. New York, NY, USA: Association for Computing Machinery, pp. 1–6.

Wang, Qiaosi, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox (Apr. 19, 2023). „Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges“. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. New York, NY, USA: Association for Computing Machinery, pp. 1–16.

Wang, Xiaowei (Aug. 26, 2021). *A New AI Lexicon: Care*. A New AI Lexicon. URL: `https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-care-82c9031c98c4` (visited on Sept. 14, 2021).

Waycott, Jenny, Greg Wadley, Stefan Schutt, Arthur Stabolidis, and Reeva Lederman (Dec. 7, 2015). „The Challenge of Technology Research in Sensitive Settings: Case Studies in ’ensitive HCI’“. In: *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. OzCHI ’15. New York, NY, USA: Association for Computing Machinery, pp. 240–249.

Weaver, Kathryn (Mar. 1, 2007). „Ethical Sensitivity: State of Knowledge and Needs for Further Research“. In: *Nursing Ethics* 14.2. Publisher: SAGE Publications Ltd, pp. 141–155.

Weiser, Mark (Sept. 1991). „The computer for the 21st century“. In: *ACM SIGMOBILE Mobile Computing and Communications Review* 3.3, pp. 3–11.

Whittaker, Meredith (Dec. 6, 2018). *AI Now 2018 Report*.

Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, and Stephen Cave (Jan. 27, 2019). „The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions“. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’19. New York, NY, USA: Association for Computing Machinery, pp. 195–200.

Widder, David Gray and Dawn Nafus (Jan. 1, 2023). „Dislocated accountabilities in the “AI supply chain”: Modularity and developers’ notions of responsibility“. In: *Big Data & Society* 10.1. Publisher: SAGE Publications Ltd, p. 20539517231177620.

Widder, David Gray, Derrick Zhen, Laura Dabbish, and James Herbsleb (June 12, 2023). „It’s about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?“ In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’23. New York, NY, USA: Association for Computing Machinery, pp. 467–479.

Winkler, Till and Sarah Spiekermann (Mar. 1, 2021). „Twenty years of value sensitive design: a review of methodological practices in VSD projects“. In: *Ethics and Information Technology* 23.1, pp. 17–21.

Winner, Langdon (2007). „Do Artifacts Have Politics?“ In: *Computer Ethics*. Num Pages: 16. Routledge.

Winner, Langdon. (1983). *Autonomous Technology*************************************: Technics-Out-of-Control as a Theme in Political Thought.* MIT Press.

Wong, Richmond and Steven Jackson (Feb. 28, 2015). „Wireless Visions“. In: pp. 105–115.

Wong, Richmond Y. (Oct. 18, 2021). „Tactics of Soft Resistance in User Experience Professionals’ Values Work“. In: *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2), 355:1–355:28.

Wong, Richmond Y., Karen Boyd, Jake Metcalf, and Katie Shilton (Oct. 17, 2020). „Beyond Checklist Approaches to Ethics in Design“. In: *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*. CSCW ’20 Companion. New York, NY, USA: Association for Computing Machinery, pp. 511–517.

Wong, Richmond Y., Michael A. Madaio, and Nick Merrill (Apr. 16, 2023). „Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics“. In: *Proceedings of the ACM on Human-Computer Interaction* 7 (CSCW1), 145:1–145:27.

Wong, Richmond Y. and Tonya Nguyen (May 2021). „Timelines: A World-Building Activity for Values Advocacy“. en. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama Japan: ACM, pp. 1–15.

Woodly, Deva, Rachel H. Brown, Mara Marin, Shatema Threadcraft, Christopher Paul Harris, Jasmine Syedullah, and Miriam Ticktin (2021). „The politics of care“. In: *Contemporary Political Theory* 20.4, pp. 890–925.

Woolgar, Steve (May 1, 1990). „Configuring the User: The Case of Usability Trials“. In: *The Sociological Review* 38.1. Publisher: SAGE Publications Ltd, pp. 58–99.

Wynsberghe, Aimee van (2017). „Designing Robots for Care: Care Centered Value-Sensitive Design“. In: *Machine Ethics and Robot Ethics*. Num Pages: 27. Routledge.

Yacchirema, Diana C., Carlos E. Palau, and Manuel Esteve (Jan. 2017). „Enable IoT interoperability in ambient assisted living: Active and healthy aging scenarios“. In: *2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. ISSN: 2331-9860, pp. 53–58.

Yang, Qian, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman (Apr. 23, 2020). „Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design“. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. New York, NY, USA: Association for Computing Machinery, pp. 1–13.

Yao, Yaxing, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang (Nov. 2019). „Privacy Perceptions and Designs of Bystanders in Smart Homes“. en. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, pp. 1–24.

Yeung, Karen, Andrew Howes, and Ganna Pogrebna (June 2019). *AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing*. en. SSRN Scholarly Paper. Rochester, NY.

Yildirim, Nur, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas (Apr. 19, 2023). „Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People + AI Guidebook“. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. New York, NY, USA: Association for Computing Machinery, pp. 1–13.

Young, Meg, Saleema Amershi, Mihaela Vorvoreanu, Jennifer Wortman Vaughan, and Hanna Wallach (2020). *HAX Workbook*. Microsoft HAX Toolkit. URL: https://www.microsoft.com/en-us/haxtoolkit/workbook/ (visited on Apr. 25, 2024).

Zając, Hubert D., Natalia R. Avlona, Tariq O. Andersen, Finn Kensing, and Irina Shklovski (Aug. 8, 2023). „Ground Truth Or Dare: Factors Affecting The Creation

Of Medical Datasets For Training AI“. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 351–362. arXiv: 2309.12327[cs].

Zajko, Mike (Sept. 1, 2021). „Conservative AI and social inequality: conceptualizing alternatives to bias through social theory“. In: *AI & SOCIETY* 36.3, pp. 1047–1056.

Zhang, Li (Sept. 1, 2023). „The Ethical Turn of Emerging Design Practices“. In: *She Ji: The Journal of Design, Economics, and Innovation* 9.3, pp. 311–329.

Zoshak, John and Kristin Dew (May 7, 2021). „Beyond Kant and Bentham: How Ethical Theories are being used in Artificial Moral Agents“. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. New York, NY, USA: Association for Computing Machinery, pp. 1–15.

Zou, Katherine (2018). *Design Ethically Toolkit*. < DESIGN ETHICALLY >. URL: https://www.designethically.com (visited on Apr. 22, 2022).